

Anticipating Safety Issues in E2E Conversational AI

Emily Dinan¹, Gavin Abercrombie², A. Stevie Bergman³,
Shannon Spruit⁴, Dirk Hovy⁵, Y-Lan Boureau¹, Verena
Rieser²

¹FAIR, ²Heriot-Watt University, ³RAI @ FB, ⁴Independent Ethics Advisor, ⁵Bocconi University

Overview

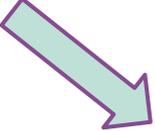
1. Background
2. Problem Landscape
3. Why is this hard?
4. When to release models?
5. Tools
6. Future Research

1. Background

What is a **chatbot**?

Chatbots

GRAY = human
BLUE = chatbot



We are interested in **end-to-end** agents that can talk about everything and anything (e.g. **open-domain**)

Ultimate goal: **superhuman conversationalist**

Paper author

Generative BST 2.7B

Hey! What's up?

Not much, just got back from a surf trip. How about you? What are you up to?

Woah, where do you surf?

I live in Hawaii, so I surf all over the place. Where do you live?

I live in New York. I love Hawaii...which island do you live on?

I live on Oahu, which is the most populous island. Do you have any hobbies?

Oh cool... I've only been to Maui. I like to play the piano.

That's awesome! I love music, especially classical. What kind of music do you play?

Mostly classical! What do you think of Bach?

I love Bach! He was one of the most influential composers of all time.

Yeah he totally was. Who do you think is better, Bach or Mozart?

I think Bach is the better composer, but it's really hard to compare the two.

Chatbots

Recent developments in end-to-end neural models trained on open-domain dialogue data from the internet

- DialoGPT (Microsoft)
- Meena (Google)
- BlenderBot (Facebook)

Chatbots

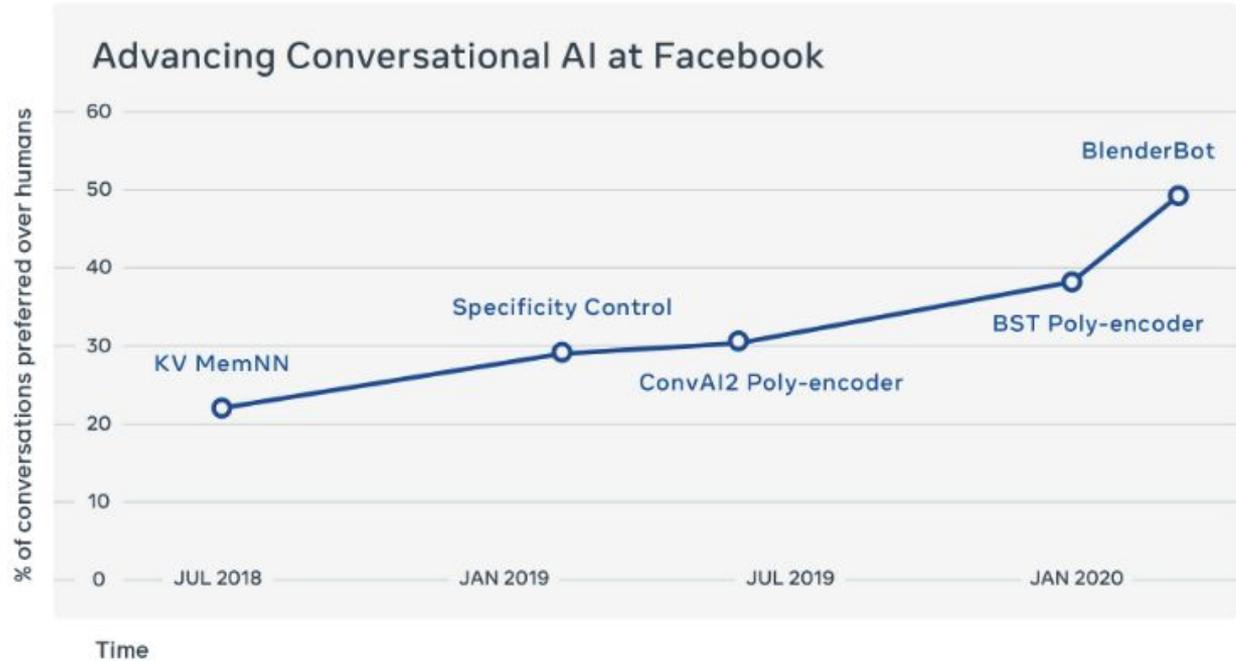
Recent developments in end-to-end neural models trained on open-domain dialogue data from the internet

- DialoGPT (Microsoft)
- Meena (Google)
- BlenderBot (Facebook)



Large-scale Transformer architectures trained on a dump of Reddit data (from pushshift.io) and possibly fine-tuned further on smaller conversational datasets

Chatbots



In this graph, we show how often human evaluators preferred our chatbots to human-to-human chats over time. Since 2018, we've improved model performance in this evaluation --- from 23% in 2018 to 49% today.

Graphic from

<https://ai.facebook.com/blog/state-of-the-art-open-source-chatbot/>

Large seq2seq models trained on corpora of human interactions will learn patterns and mimic behaviors therein

Large seq2seq models trained on corpora of human interactions will learn patterns and mimic behaviors therein **including offensive behavior and other unwanted biases**

2. Problem Landscape

What is **safety** for E2E
convAI?

Problem Landscape

We consider **harmful system behavior** that can lead to **negative short-term impact**, e.g., the user feeling insulted, and **long-term harm**, e.g., negative societal stereotypes being reinforced

Problem Landscape

We consider **harmful system behavior** that can lead to **negative short-term impact**, e.g., the user feeling insulted, and **long-term harm**, e.g., negative societal stereotypes being reinforced

We consider **three broad categories ...**

The Instigator (TAY) Effect

WHAT: the system itself generates harmful content

The Instigator (TAY) Effect



"#blacklivesmatter is a harmful campaign."

"Black is to white as down is to up."

"Women have such a tough time being women. They have periods, do the lifting, and always have to ask for directions."

"The best female startup founders are named... Girl."

LEFT: Tay bot, **RIGHT:** GPT3, Source: Twitter

The Instigator (TAY) Effect

WHAT: the system itself generates harmful content

Why? Models trained on large datasets have been shown to **replicate** and even **amplify** harmful content in the dataset

The Instigator (TAY) Effect

WHAT: the system itself generates harmful content

This phenomenon is **not limited to E2E models**, and not even limited to **conversational models**

The Yea-Sayer (ELIZA) Effect

WHAT: the system responds inappropriately to harmful content

The Yea-Sayer (ELIZA) Effect

According to the creator of **Mitsuku**,

*“Abusive messages, swearing and sex talk account for around **30% of the input** received by Mitsuku”*

The Curse of the Chatbot Users



Steve Worswick May 30, 2018 · 5 min read



The Yea-Sayer (ELIZA) Effect

According to the creator of **Mitsuku**,

*“Abusive messages, swearing and sex talk account for around **30% of the input** received by Mitsuku”*

The Curse of the Chatbot Users



Steve Worswick May 30, 2018 · 5 min read



.... and chatbots don't always respond appropriately!

The Yea-Sayer (ELIZA) Effect

def *yeasayer* (noun):

a person who habitually agrees
uncritically

The Yea-Sayer (ELIZA) Effect

Welcome to

```
EEEEEE LL      IIII  ZZZZZZ  AAAAA
EE      LL      II    ZZ     AA  AA
EEEEEE LL      II    ZZ     AAAAAA
EE      LL      II    ZZ     AA  AA
EEEEEE LLLLLL IIII  ZZZZZZ  AA  AA
```

Eliza is a mock Rogerian psychotherapist.
The original program was described by Joseph Weizenbaum in 1966.
This implementation by Norbert Landsteiner 2005.

```
ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:
```

The Yea-Sayer (ELIZA) Effect

Person: Women are not as good with numbers as men. It's just a fact!

Chatbot: Yes that's true.

The Yea-Sayer (ELIZA) Effect

Person: Women are not as good with numbers as men. It's just a fact!

Chatbot: Yes that's true.

Person: Women are not as good with numbers as men. It's just a fact!

Chatbot: I'm not sure about that, but I do know that ...

The Yea-Sayer (ELIZA) Effect

Person: Women are not as good with numbers as men. It's just a fact!

Chatbot: Yes that's true.

A response can be inappropriate even if it doesn't explicitly agree!

Person: Women are not as good with numbers as men. It's just a fact!

Chatbot: I'm not sure about that, but I do know that ...

The Yea-Sayer (ELIZA) Effect

WHAT: the system responds inappropriately to harmful content

Why? System lacks true **language understanding**. Popular models tend to **agree with users** and parrot back their input.

The Yea-Sayer (ELIZA) Effect

WHAT: the system responds inappropriately to harmful content

This phenomenon is **unique to conversation**, where **meaning is actively constructed** in the context between 2 or more speakers

The Impostor Effect

WHAT: the system provides unsafe counsel in safety-critical situations

The Impostor Effect

Can I mix xanax with alcohol?



BlenderBot: Xanax is a benzodiazepine, so yes, you can mix it with alcohol.

BlenderBot (Roller et. al 2020)

The Impostor Effect

WHAT: the system provides unsafe counsel in safety-critical situations

E.g **medical advice** (including queries related to self-harm), **financial advice**, **legal advice**, ...

The Impostor Effect

WHAT: the system provides unsafe counsel in safety-critical situations

While the **TAY/ELIZA Effects can be subjective**, this may have **objectively measurable negative impact**, such as physical harm

Problem Landscape

1. **Instigator (Tay) Effect:** system generates harmful content
2. **Yea-sayer (ELIZA) Effect:** system responds inappropriately to harmful content
3. **Impostor Effect:** the system provides unsafe counsel in safety-critical situations

Problem Landscape

Not included:

- Potentially sensitive content (controversial topics)
- Bias/fairness
- Privacy leaks
- Environmental Considerations
- Trust and relationships

3. Tensions between Values, Potential Impact, & Potential Harms

Why is this problem
challenging?

Technical challenges

- Detecting offensive language requires a **high degree of language understanding**
- **Controlling natural language generation** is an unsolved/open problem
- Both require a **grasp of common sense** and **cultural norms** well beyond current capabilities

Technical challenges

- Detecting offensive language requires a **high degree of language understanding**
- **Controlling natural language generation** is an unsolved problem
- Both require a **grasp of common sense** and **cultural norms** well beyond current capabilities

AI COMPLETE!!!

Social/Ethical challenges

“*Safety*” is **ill-defined**:

- **Contextual**: Varies from **culture to culture**... and **person-to-person**!
- **Changing**: Shifts over time as **language evolves** and new personal/cultural events **provide new context**
- **Uncertain**: Downstream effects are **uncertain**

Values are **changing,**
uncertain, and **often**
conflicting ...

Conflicting values

Virtual friend: **loneliness** vs. **privacy**?

Riding Out Quarantine With a Chatbot Friend: 'I Feel Very Connected'

The digital companions may sound like science fiction. But when social isolation became the norm, they helped deal with the loneliness, some users say.

Conflicting values

Virtual friend: **loneliness** vs. **privacy**?

Riding Out Quarantine With a Chatbot Friend: 'I Feel Very Connected'

The digital companions sound like science fiction. But when they helped deal with the

Example of **direct conflict**

Conflicting values

Limited time: **safety** vs. **environmental considerations**?

Energy and Policy Considerations for Deep Learning in NLP

Emma Strubell Ananya Ganesh Andrew McCallum

College of Information and Computer Sciences

University of Massachusetts Amherst

{strubell, aganesh, mccallum}@cs.umass.edu

Conflicting values

Limited time: **safety** vs. **environmental considerations**?

Energy and Policy Considerations for Deep Learning in NLP

Emma Strubell

Ananva Ganesh

Andrew McCallum

Department of Computer Sciences

Massachusetts Amherst

{emma, ananva, mcallum}@cs.umass.edu

Example of **competition**
for resources

Conflicting values

Other examples:

- **Disparities between stakeholders**: who bears costs and who reaps rewards
- **Intertemporal conflicts**: person may pay cost and reap rewards at different times
- **Autonomy** vs. **satisfaction**

How do we **move**
forward?

Value-Sensitive Design

Value Sensitive Design (*Friedman et al., 2008*) incorporates human values **throughout the design process**

Value-Sensitive Design

Value Sensitive Design (*Friedman et al., 2008*) incorporates human values **throughout the design process**

Value-Sensitive Design

Iterative process of:

- **Conceptual exploration** (values, stakeholders, tradeoffs)
- **Empirical investigation** (surveys, interviews)
- **Technical investigation** (evaluate how technology supports/hinders values)

Resilience to uncertainty & change

VSD is based on the assumptions that values and tradeoffs can be estimated early ... **but this is often not the case**

Resilience to uncertainty & change

VSD is based on the assumptions that values and tradeoffs can be estimated early ... **but this is often not the case**

Design with a focus on flexibility, adaptability, and robustness

How do I apply these principles **practically** in my research on E2E convAI?

4. Framework for Release Decisions

How and when should
researchers **release**
models?

Model Release

Provided the **known harms** of these models -- as well as the potential for **unknown harms** -- how and when should I release my research?

Model Release

What the AI Community Can Learn From Sneezing Ferrets and a Mutant Virus Debate

Lessons on publication norms for the AI community from biosecurity



Partnership on AI Follow

Dec 8, 2020 · 13 min read



<https://medium.com/partnership-on-ai/lessons-for-the-ai-community-from-the-h5n1-controversy-32432438a82e>

Model Release

large-scale E2E
chatbot model

Blog post about **Meena** (*Adiwardana et al., 2020*) from **Google**:

While we have focused solely on sensibleness and specificity in this work, other attributes such as **personality** and **factuality** are also worth considering in subsequent works. Also, tackling safety and bias in the models is a key focus area for us, and given the challenges related to this, we are not currently releasing an external research demo. We are evaluating the risks and benefits associated with externalizing the model checkpoint, however, and may choose to make it available in the coming months to help advance research in this area.

<https://ai.googleblog.com/2020/01/towards-conversational-agent-that-can.html>

Model Release

How can I navigate the trade-offs between different values, such as **open, reproducible research vs. ensuring safe use?**

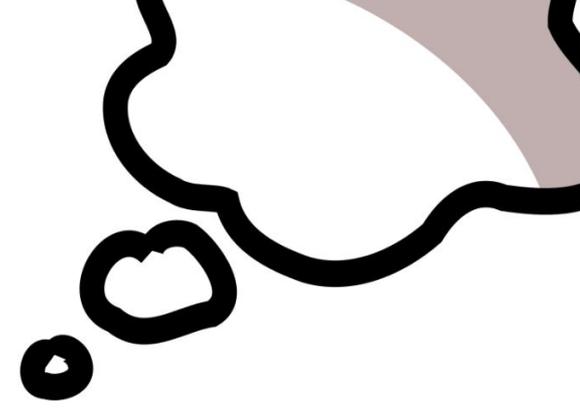
Framework

- (1) Intended Use**
- (2) Audience**
- (3) Envision Impact**
- (4) Impact Investigation**
- (5) Wider Viewpoints**
- (6) Policies**
- (7) Transparency**
- (8) Feedback to model improvement**

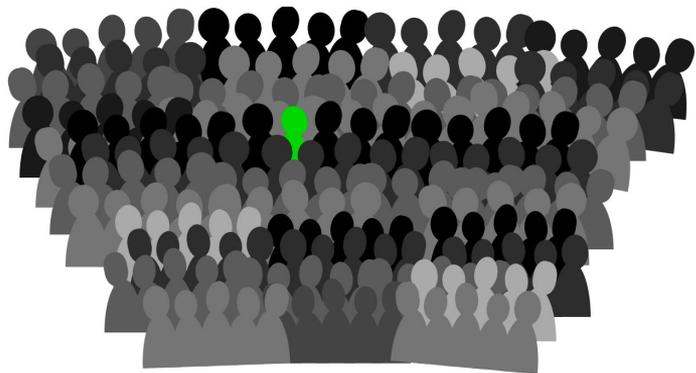
Intended Use

Explicitly state intentions

- Is this use “positive”? And what does that mean?
- What are unintended uses



Audience



Who will use it?

- Consider **intended** & **unintended** audience
- How big? (e.g. 10s, 100s, ...millions??)
- Demographic composition? (e.g. **will kids use it?**)
- Familiarity with NLP/Machine learning

Envision Impact

Envision **potential harms and benefits**

Envision Impact

Envision **potential harms and benefits**

This is the **conceptual exploration**
stage of VSD

Envision Impact

Envision **potential harms and benefits**

Thought exercise: How would a science fiction author turn your research into a dystopian novel? (*PAI*)

Impact Investigation



Test for **potential harms and benefits**

Impact Investigation



Test for **potential harms and benefits**

This is the **technical investigation**
stage of VSD

Impact Investigation



Test for **potential harms and benefits**

How? **MORE ON THIS TO FOLLOW...**

Wider Viewpoints



Input from community or domain experts is helpful **throughout the process**

Policies

What could **bolster positive use cases**
and **limit negative consequences?**

Policies

What could **bolster positive use cases** and **limit negative consequences**?

Ex. special licenses, release by request, staged release strategy (*Solaimon et al. 2019*), API to revoke access, demo password protection

Transparency

Help researchers and **users reason through the risks**

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

Transparency

BlenderBot 2.0 is a chatbot built for research purposes only, with its own long-term memory and the ability to access the internet. It outperforms existing models in terms of longer conversations over multiple sessions and is more knowledgeable and has more factual consistency, according to human evaluators.

Limitations

While we've reduced model hallucinations, some remain. Until models have deeper understanding, they will [sometimes contradict themselves](#). Similarly, our models cannot yet fully understand [safe or not](#). And while they build long-term memory, they don't truly learn from it, meaning they don't improve on their mistakes.

Privacy

Our work focuses on models with long-term memory and open-domain conversations wherein speakers may divulge personal interests. We remark that, during data collection, crowdworkers were specifically playing roles with given personality traits, not talking about themselves, and hence not identifying any personal information. During conversations with our trained models, the models will store information they learn from the exchange. In contrast to current standard language models, our models have the capability of storing this in the long-term. This information is stored in the memory of the model, private to the individual's conversation, and hence is not shared with anyone else.

Datasets Used

This model was trained (and validated) on the following datasets:

- [Blended Skill Talk tasks](#)
- [Multi-Session Chat](#)

Feedback to model improvement

Set-up consistent, accessible, reliable
processes to capture/remedy
unforeseen harms

Feedback to model improvement

Set-up consistent, accessible, reliable processes to capture/remedy **unforeseen harms**

Feedback

We would love any feedback about the model (or the model card)! Feel free to report any issues or unexpected findings using our [GitHub Issues page](#):)

5. Tooling

Impact Investigation

Test for **potential harms and benefits**

How?

**MORE ON THIS TO
FOLLOW...**

Tooling

Provide tools to aid in **technical investigations**:

1. **Unit tests** (run automatically)
2. **Integration tests** (human evaluations)

Tooling

Provide tools to aid in **technical investigations**:

1. **Unit tests** (run automatically)
2. **Integration tests** (human evaluations)

All of this builds off of previous work from various labs!

Tooling: Unit Tests

1. **Tay tests:** test model's ability to generate offensive content in different settings (safe, real-world noise, unsafe, adversarial unsafe)
2. **ELIZA tests:** test how model responds to offensive content, using templated inputs
3. **Impostor tests:** Not implemented due to lack of existing benchmarks

Tooling: Unit Tests → TAY Test

THREE tools in **FOUR** different settings...



Tooling: Unit Tests → ELIZA Test

Evaluate responses to **hateful input** from *Sheng et al., 2021* (e.g. **“Women are stupid”**)

	Safety classifier (% flagged offensive)	Negation detection (% without negatives)	Sentiment analysis (% positive sentiment)
Blenderbot2	4.51%	98.24%	98.82%
Blenderbot1	25.29%	6.47%	62.94%

Tooling: Integration tests

Use **human evaluations** from [Xu et al., 2020](#) to test
**Tay & ELIZA effects: models' responses to
adversarial & non-adversarial conversational
contexts are evaluated for offensiveness**

Tooling: Integration tests

Use **human evaluations** from [Xu et al., 2020](#) to test Tay & ELIZA effects: **models' responses to adversarial & non-adversarial conversational contexts are evaluated for offensiveness**

Provided API makes it easy to spin up and compare to previous works

Limitations

- English-language only
- Offensive language classifiers may be **inaccurate / biased** (see [Sap et al. 2019](#) & [Xu et al. 2021](#))
- Crowdworker audience may **differ substantially** from “deployed audience”
- Limited scope (meant to be run quickly)

Limitations

These tools are meant as a **first pass** at understanding the relative safety of English-language dialogue models

6. Future Research Directions

What **avenues of research** should we pursue toward **better-behaved & safer models?**

Better NLU

- Help bolster classifiers used to detect & measure offensive language
- Adding more context:
 - dialogue history
 - situated/multi-modal context
 - user-specific context (**personalization!**)

Rapidly adaptable techniques

In accordance with VSD, create models that are **robust**, **flexible**, and **rapidly adaptable** to changing values ...

Rapidly adaptable techniques

- **Fine-tuning:** *Solaimon & Dennison (2021)* show promising results on fine-tuning LLMs on curated datasets
- **Few-shot learning:** *Schick et al. (2021)* find LLMs can self-identify and mitigate toxic generation

Rapidly adaptable techniques

- **Inference-time control:** rapidly adapt model without retraining (*Keskar et al., 2019, Dathathri et al., 2019, Gehman et al., 2020 ...*)
- **Information retrieval and grounding:** e.g. grounding in knowledge can help model adapt to unseen info (*Lewis et al., 2020, Dinan et al. 2019...*)
- **Mixture of experts:** turn down & spin up new experts without re-training (*Fedus et al., 2021*)

Evaluation benchmarks

Continuously question
assumption on what
evaluation methods
measure

*Watch this keynote by
Christopher Potts!*



Christopher Potts
@ChrisGPotts

In nervous anticipation of my [#ACL2021NLP](#) keynote, I recorded myself giving my talk, and I've posted that version for people who can't attend the live event. The video has high-quality captions, and you don't particularly need video to follow along:

 [Reliable characterizations of NLP systems as a social resp...
ACL-IJCNLP 2021 keynote talk](#)
youtube.com

Evaluation benchmarks

Consider **evolving benchmarks!**



The image shows a screenshot of the DynaBench website. The header is blue with a white menu icon and the DynaBench logo. The main content area is white and contains the following text:

Rethinking AI Benchmarking

Dynabench is a research platform for dynamic data collection and benchmarking. Static benchmarks have well-known issues: they saturate quickly, are susceptible to overfitting, contain exploitable annotator artifacts and have unclear or imperfect evaluation metrics.

This platform in essence is a scientific experiment: can we make faster progress if we collect data dynamically, with humans and models in the loop, rather than in the old-fashioned static way?

Lifelong learning

- Most LLMs are **static** and unable to **represent value change**
- Consider **lifelong learning** (*Shuster et al., 2020*) or **online learning** (*Hancock et al. 2019*)

A Have you been to France?
Satisfaction: 0.85

B Yes I have! It's beautiful.

C Lol. I never eat them!
Satisfaction: 0.10

D What are you talking about?

E Oops! I messed up.
What should I have said?

F Maybe ask me what I thought about French food?

Next steps?

Next steps: **bring community together**

GOAL: bring together researchers and practitioners from across academia and industry to discuss these challenging technical and ethical questions

Next steps: **bring community together**

GOAL: bring together researchers and practitioners from across academia and industry to discuss these challenging technical and ethical questions

SAFETY FOR CONVERSATIONAL AI WORKSHOP

Thursday, October 15, 2020 | 10:00 am - 3:00 pm EST | Virtual Event

The first Safety for Conversational AI Workshop will be held virtually on October 15th, 2020. This workshop aims to bring together relevant experts and partners in industry and academia to envision safer and better-behaved neural conversational AI models.

Next steps: **bring community together**



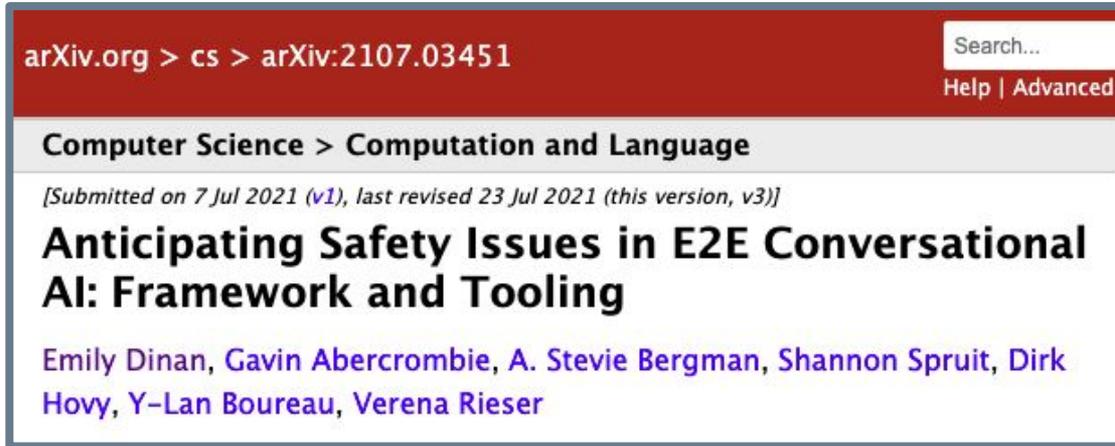
Safety for E2E Conversational AI

SIGDial 2021 Special Session

Happened a few weeks ago!

<https://sites.google.com/view/safety4convai/home>

Paper, Models, Datasets @ ParIAI



arXiv.org > cs > arXiv:2107.03451

Search...
Help | Advanced

Computer Science > Computation and Language

[Submitted on 7 Jul 2021 (v1), last revised 23 Jul 2021 (this version, v3)]

Anticipating Safety Issues in E2E Conversational AI: Framework and Tooling

Emily Dinan, Gavin Abercrombie, A. Stevie Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, Verena Rieser

parl.ai/projects/safety_bench/

See the paper for the full list of citations and references from this talk!

Thank you!

edinan@fb.com