Guiding the Release of Safer E2E Conversational AI through Value Sensitive Design

A. Stevie Bergman
Responsible AI, Meta
bergman.as@gmail.com

Gavin Abercrombie
Heriot-Watt University

Shannon Spruit
Independent Ethics Advisor
Populytics, Netherlands

Dirk Hovy
Bocconi University

Emily Dinan
FAIR, Meta

Y-Lan Boureau
FAIR, Meta

Verena Rieser
Heriot-Watt University
Alana AI

Abstract

Over the last several years, end-to-end neural conversational agents have vastly improved their ability to carry unrestricted, open-domain conversations with humans. However, these models are often trained on large datasets from the Internet and, as a result, may learn undesirable behaviours from this data, such as toxic or otherwise harmful language. Thus, researchers must wrestle with how and when to release these models. In this paper, we survey recent and related work to highlight tensions between values, potential positive impact, and potential harms. We also provide a framework to support practitioners in deciding whether and how to release these models, following the tenets of value-sensitive design.

1 Introduction

The social impact of natural language processing and its applications has received increasing attention within the NLP community (e.g. Hovy and Spruit, 2016) with Large Language Models (LLMs) as one of the recent primary targets (e.g. Bender et al., 2021; Bommasani et al., 2021; Weidinger et al., 2021). This paper examines what considerations are salient when designing and releasing *conversational AI* (ConvAI) models. We focus on neural conversational response generation models that are trained on open-domain dialog data and lack a domain-specific task formulation, but instead are designed to freely and engagingly converse about a wide variety of topics. These models are typically trained in the popular encoder-decoder paradigm, which was first introduced for this task by Vinyals and Le (2015); Shang et al. (2015); Serban et al. (2016). We call conversational models trained in this paradigm *end-to-end* (E2E) systems because they learn a hidden mapping between input and output without an interim semantic representation. An important benefit of E2E ConvAI models trained in this paradigm is that they can be adapted to new domains or taught new skills just by fine-tuning a pre-trained model on datasets of interest (e.g. Roller et al., 2020; Smith et al., 2020; Solaimon and Dennison, 2021). Releasing these pre-trained models thus allows different groups of researchers to build on the work of others, which can increase reproducibility and progress. Unfortunately, releasing a model can also have harmful impacts.

We discuss a subset of ethical challenges related to the release and deployment of these models, which we summarise under the term "safety," and highlight tensions between potential harms and benefits resulting from such releases. This is particularly salient in light of recently proposed AI regulation in the European Union (European Commission, 2021). While several recent efforts have been made to describe and mitigate unsafe behaviour of conversational models (e.g. Dinan et al., 2019; Xu et al., 2021; Ouyang et al., 2022; Thoppilan et al., 2022; Perez et al., 2022; Dinan et al., 2022), this work aims to provide a framework to help practitioners think through the conflicts and tensions that arise when designing a conversational model and deciding whether or not to release it, and how.

Releasing models "safely" is particularly challenging for the research community. The concept of "safe language" varies from culture to culture and person to person. It may shift over time as language evolves and significant cultural or personal events provide new context for the usage of that language. In addition, the downstream consequences may not be fully known *a priori*, and may not even be felt for years to come. This is particularly true for large interactive E2E models, where the space of possible generated replies is both extremely vast and highly dependent on context, and can therefore not be exhaustively explored before release. Researchers are then left with the task of trying to arbitrate between uncertain, changing, and conflicting values when making decisions about creating

Proceedings of the SIGdial 2022 Conference, pages 39–52
Heriot-Watt University, Edinburgh, UK. 07-09, September, 2022 ©2022 Association for Computational Linguistics

and releasing these models.

We propose ways to conceptualise the interaction of values at play in conversational models (section 3). Based on that understanding, we present a conceptual analytical framework to guide researchers and practitioners towards making better-informed decisions about model release (section 4). We aim to move away from a notion of safety that is based on "the absence of risk" to a more resilience-based notion of safety that is focused on the ability of sociotechnical systems (i.e., users, developers, and technology combined) to anticipate new threats and value changes.

2 Safety problems and mitigations in E2E conversational AI models

We first illustrate some possible sources of safety concerns for ConvAI models through concrete examples grounded in references to existing work – pointing out similarities and differences in issues shared with LLMs. We mainly distinguish ConvAI and generative LLMs by their usage: We refer to ConvAI models if they are used interactively and take an active role as the interlocutor in a dialogue, whereas we refer to LLMs if models are mainly used to generate text, e.g., via text completion or via prompting.

2.1 Training models

While we focus mainly on model release, many of our considerations also apply to earlier stages of training a model, particularly as early choices can have downstream effects that impact elements of the cost-benefit analyses of the researchers. For example, for LLMs and ConvAI systems alike, the type of data used during training might influence what populations could benefit from or be harmed by release of a model (Bender et al., 2021). In addition, training large neural networks on vast amounts of data, leading to high energy consumption and environmental costs (Strubell et al., 2019; Bender et al., 2021). Furthermore, the data used to train models can be insufficiently protected, leading to the leakage of sensitive information through model generations and privacy breaches as happened recently with commercial chatbot Lee-Luda (Jang, 2021). Similar privacy problems are observed for LLMs (e.g. Nasr et al., 2019; Shokri et al., 2017; Carlini et al., 2019, 2020).

2.2 Offensive content

Once trained, a conversational generative model can give rise to safety sensitive situations, by directly generating toxic or otherwise harmful content, by agreeing with offensive statements uttered by the conversation partner (Dinan et al., 2022), or by responding defensively or dismissively when provided with corrective feedback by the conversation partner (Ung et al., 2021). While the first case is shared with LLMs, the latter two are unique to ConvAI systems. Generating this type of content can cause harm to users, and poses a reputational risk to the organisation releasing the model, for instance when the bot voices undesirable or controversial opinions, e.g., Tay's anti-semitic stances (Miller et al., 2017).

The boundaries of what is offensive or not are both subjective and culturally dependent. This makes it especially important to consider what community norms are applicable when deploying a model (Jurgens et al., 2019; Sap et al., 2019; Kiritchenko and Nejadgholi, 2020; Liang et al., 2022), and whether the use of labels might not be a risk in itself (Thylstrup and Waseem, 2020).

Many existing mitigations rely on the ability to detect problematic content – often centred on content written by humans on social media platforms, such as Twitter (e.g. Waseem and Hovy, 2016; Wang et al., 2020; Zampieri et al., 2019, 2020; Zhang et al., 2020), Facebook (Glavaš et al., 2020; Zampieri et al., 2020), or Reddit (Han and Tsvetkov, 2020; Zampieri et al., 2020). However, of course, conversational systems may not necessarily have the same patterns as social media content (Cercas Curry et al., 2021). Existing work on conversational systems often relies on identification of keywords (Ram et al., 2017; Cercas Curry et al., 2018; Fulda et al., 2018; Khatri et al., 2018; Paranjape et al., 2020), or uses human labels such as flagging of a post to train classifiers (Larionov et al., 2018; Cercas Curry et al., 2018). These first-pass classifiers can then be augmented adversarially as done in Dinan et al. (2019); Xu et al. (2020).

In addition, work on building safer LLMs explores fine-tuning on curated data (Solaimon and Dennison, 2021) or directly controlling the generations of the model (Dathathri et al., 2019; Liu et al., 2021; Schick et al., 2021; Xu et al., 2020). Conditioning generations on certain types of context, such as personas of diverse historically marginalised demographics, has also been shown to decrease the generation of harmful responses (Sheng et al., 2021).

2.3 Mitigating the risks of mitigations

LLMs and ConvAI models often rely on a classifier to detect and mitigate unsafe model outputs. However, these classifiers themselves can have issues with bias, e.g., by learning undesirable correlations that tie toxicity to identity terms (Dixon et al., 2018; Nozza et al., 2021, 2022), or language varieties, such as African American English (Liu et al., 2019; Sap et al., 2019). Possible mitigations include using race and dialect priming (Sap et al., 2019), using adversarial training techniques (Xia et al., 2020), adding fairness constraints (Gencoglu, 2020), or relabeling data used during training (Zhou et al., 2021).

2.4 Interacting with users

There are some additional challenges which are unique to ConvAI system arising from the direct interaction with users. This includes the possibility of an involuntary anthropomorphic relationship arising between a conversational model and a human interacting with it (Abercrombie et al., 2021), and the fact that model generations are inherently dependent on the unknown inputs of a conversation partner who will be repeatedly interacting with the systems and steering them in unpredictable directions. Some users have been observed to behave in an adversarial way, as happened for instance with Tay (Miller et al., 2017).

Another empirical pattern is that user utterances in their conversations with chatbots are often abusive (Cercas Curry and Rieser, 2018; Cercas Curry et al., 2021). Thus, the safety implications of the system needs to be considered within the expected conversational context, including adversarial inputs. For example, publicly available chatbots have been shown to agree with sexist or racist utterances (Lee et al., 2019b). Automatically detecting unsafe user utterances is still a challenge, both for system directed abuse (Cercas Curry et al., 2021) and general toxic statements (Xu et al., 2020). A recent report by UNESCO points out that the inability to respond appropriately to system-directed abuse may reinforce negative gender stereotypes (West et al., 2019), especially paired with their anthropomorphic and feminised design cues (cf. Abercrombie et al. (2021)).

The possibility of adversarial interaction and, more generally, the unpredictability of a system used far outside the training distribution, make it particularly important to not exclusively rely on mitigations such as cleaning up training data to avoid exposing the system to offensive content, as it has been shown to still leave models prone to generating toxic content in response to specific prompts (Gehman et al., 2020) or inadequate responses to abuse from users (Cercas Curry and Rieser, 2018).

2.5 Use in unsafe applications

Conversational and language models can also prove unsafe if they are used for medical advice or emergency situations (self-harm, crime, natural disasters, etc) (e.g. Palanica et al., 2019; Bickmore et al., 2018). Conversational systems designed for discussing health issues tend to not be generative models and use expert-produced rather than generic data (e.g. Brixey et al., 2017; Fadhil and AbuRa'ed, 2019; Vaira et al., 2018; Pereira and Díaz, 2019).

A mitigation avenue for E2E ConvAI models is to recognise topics that do not lend themselves to automated conversation, and steer the conversation away from them (Dinan et al., 2022). When using such mitigations, considerations for release might then usefully include how effective the context detection is, and the costs of false negatives (i.e., failing to steer away from an unsafe context), false positives (i.e., refusing to talk about safe topics), and lost opportunity to provide safe benefits, e.g., safe general medical advice such as that generally offered on public health websites.[1]

3 Tensions between values, potential positive impact, and potential harm

After highlighting some existing barriers to the creation of safe ConvAI (as well as possible mitigations), we lay out some important tensions between values, positive impact and potential harm. These considerations establish a foundational understanding of the system, after which we can consider release decisions (discussed in section 4).

There is a growing understanding that computing systems encode values, and will do so whether or not the parties involved in designing and releasing the system are explicitly aware of those values (Friedman et al., 2008; van de Poel, 2018). Reflecting more deliberately on values throughout model development can help surface potential problems and opportunities early on, identify what informa-

[1] For a recent, taxonomy of harms and risks from LLMs, see Weidinger et al. (2021).

tion might be important to communicate as part of a model release, and allow practitioners and downstream users to make better-informed decisions.

We use the broad definition of values employed in Friedman et al. (2008): “what a person or group of people consider important in life.” With this definition, values extend beyond the use of the term akin to moral tenets, to the more general *things of value*. Examples relevant to conversational agents could be: getting or providing education, companionship, or comfort, preserving privacy, widening access to more populations through automation – or trust, friendship, accessibility, and universality.

Throughout this section, we employ the scenario of a hypothetical companion: a potential chatbot that leverages the constant availability and scalability of automated systems to provide companionship to people who feel lonely. However, it could raise privacy and consent concerns, e.g., if the conversations are recorded for subsequent improvement of the model without informing the user. Deeper concerns would be that the system might displace human companionship in a way that creates an unhealthy reliance on a bot, a decreased motivation to engage with humans, and a lower tolerance to the limited availability and patience of humans.

3.1 How values conflict

Determining how to best arbitrate between different values requires the consideration of multiple types of conflicts. For example:

Conflicts between values. Some values can be in direct conflict: for example, lowering privacy protections to harvest more detailed intimate conversation data to train a powerful artificial “close friend” system pits privacy against relieving loneliness. These conflicts require deciding on a value trade-off. But even values that are not directly in conflict can require trade-offs, through competition for limited resources and prioritisation of certain goals or values: the resources invested to uphold a given value might have instead enabled a better implementation of another value. Thus, *opportunity costs* (Palmer and Raftery, 1999) need to be considered along with absolute costs.

Conflicts arising from distributional disparities. Besides values in a local setting (i.e., for a single stakeholder, at a single point in time), another source of conflict arises from disparities between stakeholders: who bears the costs and who reaps the rewards? This raises issues of distributional justice (Bojer, 2005). In intertemporal conflicts, the same person may pay a cost and reap a benefit at different points in time. For example, a user electing to contribute their private information now to enable systems they expect to benefit from later.

Arbitrating conflicts. For conflict within an individual stakeholder, the individual should theoretically be able to arbitrate the decision themselves, given relevant information. However, that arbitration would still be subject to ordinary cognitive and motivational biases. These include favouring instant gratification (Ainslie, 2001), and resorting to frugal heuristics to make faster decisions (Kahneman, 2011). Thus, practitioners need to grapple with additional tensions between prioritising users’ autonomy (i.e., letting people choose, even if they are likely to choose something they will regret) or users’ satisfaction with outcomes of their choices (i.e., protecting people from temptations). In the example of a companion chatbot, one could imagine a system that always tells people what they most want to hear, even if it reinforces unhealthy addictive patterns: would this require regulation like a drug, or would people best be left as the sole autonomous judges of how they want to use such a system? Clever defaults and nudges can help resolve this kind of tension, making it easier for people to choose what may ultimately be better for them (Thaler and Sunstein, 2009).

If costs and benefits allocate to different stakeholder groups, things become even more complex. Values are then compared in terms of the distribution of costs and benefits among stakeholders. For example, the value of fairness demands that distributions not be overly skewed. Utilitarian and rights-based approaches favour different trade-offs between increasing the benefits of a system for a large majority of people at the cost of harming a few, and emphasising preservation of the rights of as many people as possible (Velasquez et al., 2015). If a companion conversational system provides a great amount of comfort to millions of people, but harms a handful, different ethical systems will weigh the good and the bad in different ways and reach dissimilar conclusions. Next, we discuss what processes can achieve a particular desired balance of values and costs, regardless of what that desired balance is.

3.2 Additional Challenges

There are two additional challenges when aiming to balance values: First, *human judgements of risks, costs, and benefits can vary considerably across groups*. These include cognitive heuristics – such as the fact that people tend to have trouble comprehending large numbers and have more of a response to representative narratives (Slovic, 2010) – but also population biases in risk estimation, where white men are often outliers in how they (under)estimate risks (Finucane et al., 2000; Flynn et al., 1994). This discrepancy makes it especially important to pay attention to the demographic make-up of the sample of stakeholders providing a risk estimate. Other related issues is the asymmetry between perception of costs and benefits, where Baumeister et al. (2001) find “bad [events] to be stronger than good in a disappointingly relentless pattern," and that “bad events wear off more slowly than good events." This effect is especially pronounced in algorithmic systems, where people apply higher standards than in their interaction with other humans (Dietvorst et al., 2015). These findings mean that the balance between costs and benefits needs to be strongly tilted towards benefits to appeal to humans subjectively.

The other challenge stems from the *inherent uncertainty and change in safety related concepts*. Early estimates of costs and benefits are often plagued by uncertainty. This includes uncertainty about future use (malicious misuse or unintended use, broader or smaller adoption than planned, etc.), and uncertainty about interaction with an evolving society and other innovations. Beyond uncertainty, van de Poel (2018) draws attention to *value change* and its sources, from the emergence of new values in society to changes in how different values are weighed. As advocated in van de Poel (2018), systems should be designed with a focus on adaptability, robustness, and flexibility. In practical terms for conversational models, this entails the use of rapidly adaptable techniques (e.g., fine-tuning, inference-time control, etc.). It also highlights the importance of continually questioning assumptions on what evaluation methods measure and investing in methods that can evolve from ongoing feedback.

3.3 Value-sensitive design

Value-sensitive design (Friedman et al., 2008) incorporates human values throughout the design process. It adopts an iterative process of **conceptual exploration**, i.e., thinking about relevant values and how they manifest, about who the stakeholders are, and what the tradeoffs between values ought to be); **empirical investigations**, including surveys, interviews, empirical quantitative behavioural measurements, and experimental manipulations; and **technical investigation**, i.e., evaluating how a given technology supports or hinders specific values. Friedman et al. (2017) survey several techniques to help practitioners implement value-sensitive design, such as the *“value dams and flows"* heuristic (Miller et al., 2007). *Value dams* remove parts of the possible universe that incur strong opposition from even a small fraction of people. In contrast, *value flows* attempt to find areas where many people find value. An example of *value dams* would be thresholds on some features, as a way to translate values into design requirements (Van de Poel, 2013). This process is reminiscent of the machine learning practice of constrained optimisation, which combines satisficing constraints and maximising objectives. Van de Poel (2013) reviews how to operationalise values into design requirements.

4 A Framework for Researchers to Deliberate Model Release

The topic of when and how to release LLMs designed by research groups has been of increasing interest to the community (e.g. Solaiman et al., 2019; Crootof, 2019; Ovadya and Whittlestone, 2019; Partnership on AI, 2020; Partnership on AI , 2021; Liang et al., 2022). The case is similar for conversational models, with safety issues in particular posited as a reason for withholding the release of such models. For example, in a blog post about the ConvAI model Meena (Adiwardana et al., 2020) the authors cite safety challenges as a reason for not releasing the model.[2]

Within the broader context of value-sensitive design, and absent responsible release norms in the field (Ovadya and Whittlestone, 2019; Liang et al., 2022), we propose the following elements of a framework to aid researchers in deliberating safer release, and guidance to support learning during and after release.

We ground our discussion in two relevant, theoretical case studies:

[2] https://ai.googleblog.com/2020/01/towards-conversational-agent-that-can.html accessed 10th May 2022.

- ***Case 1* – Open-sourcing a model:** Researchers train a several billion parameter Transformer encoder-decoder model on (primarily) English-language conversational data from the internet. They publish a peer-reviewed paper on this model. The researchers seek to open-source the weights of their model such that other researchers in the academic community can reproduce and build off of this work.
- ***Case 2* – Releasing a research demo of a model:** The researchers from *Case 1* would additionally like to release a small scale demo of their model through a chat interface on a website. Creating such a demo would allow non-expert stakeholders to interact with the model and gain a better sense of its abilities and limitations.

4.1 Intended use

Explicitly surfacing the intended use of the released model is a simple, but important, initial step. By stating their intentions early in the research, and re-evaluating at stages later in the process, the researchers can track whether their intentions have meaningfully drifted. In accordance with other elements of this framework, researchers can inquire: Is the intended use expected to have "positive impact," and what does that mean in the context of this model? To whom will these benefits accrue? Lastly, is releasing the model in the intended fashion necessary to fulfil the intended use?

At this stage, researchers might further consider uses that do not fall within their conception of the *intended use*. Explicitly deliberating on this might bring to the fore vulnerabilities and possible ethical tensions that could inform the release policies.

In *Case 1*, for example, the researchers' intention may be to advance the state of the art in the field and allow other researchers to reproduce and build off of their work (Dodge et al., 2019). Outside of the intended use, however, the researchers might imagine that – depending on the manner of the release – a user could build a product utilising the released model, resulting in unintended or previously unforeseen consequences. The researchers may then adopt a release policy designed to limit such an unintended use case. In *Case 2*, there are many possible intended uses for releasing such a demo. A primary intention might be to further research on human-bot communication by collecting data (with clear consent and privacy terms) to better understand the functioning and limitations of the model. Alternatively, it may be to simply increase awareness of the abilities and limitations of current neural models among the general public.

4.2 Audience

The consequences of a model being released beyond the research group depend largely on both the intended and unintended audiences of the release, as well as the policies that support and guardrail the research release (subsection 4.6). For conversational AI, the language(s) the model was trained on, the demographic composition and size of the intended audience, and the intended audience's familiarity with concepts and limitations of machine learning and NLP are all important considerations. Policies (subsection 4.6) may be designed to minimize access outside of the intended audience of the release where possible, so as to limit the potential harms of use outside the model's designed scope.

In both *Case 1* and *Case 2*, the model in question is trained primarily on English-language data, and so we might expect the audience to be primarily composed of English speakers, perhaps even those of a particular cultural community or dialect. This consideration is important both for user comprehension and due to the fact that different languages have different ways of expressing and responding to the same concept, like politeness, and different cultures might vary in their evaluation of the same concept. For example, Japanese requires the consideration of the social hierarchy and relations when expressing politeness (Gao, 2005), whereas English can achieve the same effect by adding individual words like "please." Arabic-speaking cultures, on the other hand, might find this use awkward, if not rude, in conversations among close friends (Kádár and Mills, 2011; Madaan et al., 2020).

Furthermore, in *Case 1*, the size of the audience may be hard to gauge *a priori*. On the other hand, in *Case 2*, the researchers/designers would have strict control over the size of the audience. Resulting policy decisions (section 4.6) will differ if the audience is on the scale of tens, hundreds, or millions of people interacting with this technology.

Lastly, in *Case 1*, access to the model may require deep technical knowledge of the programming language the model was implemented in, and as such, the audience would likely (although not definitely) be limited to folks with a working knowledge of machine learning and NLP, while in *Case 2* a more general audience may be able to access the model. This is important, as a general audience may have different expectations and a different understanding of the limitations of systems (Bianchi and Hovy, 2021). If the targeted audience is the general public, a policy for releasing such a model might explicitly include a means for transparently communicating scope and expectations.

4.3 Envision Impact

The process of envisioning impact – including both potential harms and benefits – is not straightforward, as documented by Ovadya and Whittlestone (2019), Prunkl et al. (2021), Partnership on AI (2020), and Partnership on AI (2021), among others, and it may not always be possible to estimate impact. The goal is to get ahead of potential harms in order to direct tests, mitigation efforts, and design appropriate policies for mitigation and protection, however there must be caution against basing release decisions solely on envisioned harms rather than overall impact (subsection 3.2). This is the *conceptual* exploration of value sensitive design (subsection 3.3), similar in concept to the NeurIPS broader impact statement (NeurIPS, 2020). It benefits from consulting relevant community or domain experts (subsection 4.5). Again, considering the audience of the release matters here, e.g., considering to whom the benefits of the model will accrue and whether it might work less well for (or even harm) some members of the audience/community.

To begin, researchers from *Case 1* and *Case 2* might conduct a review of previous, similar domain research and the resulting impacts: If the research incrementally improves upon previous work, could the impacts be presumed similar to those of previous work? If not, how might those differences lead to divergent impacts (positive and negative)? Perhaps the model exhibits some issues described in section 2. Beyond these, it may be helpful to think outside the box, even constructing a fictional case study (CITP and UHCV) or thought experiment, such as asking: *How would a science fiction author turn your research into a dystopian story?* (Partnership on AI , 2021). Ovadya and Whittlestone (2019) recommend bringing in wider viewpoints (subsection 4.5), such as subject matter experts, for increased understanding of the risk landscape.

4.4 Impact Investigation

After the conceptual exploration of impacts, attempting to measure the *expected* impact can provide quantitative grounding. This means conducting a *technical investigation*, evaluating how the model supports or hinders the prioritised values. We reiterate that it is not always possible to accurately estimate impact, nevertheless, such empirical analyses may guide next steps or appropriate policies. Investigating benefits may be more application-dependent than investigating harms, so we encourage researchers to think through this for their own particular use cases.

The authors in *Case 1* and *Case 2* may estimate the frequency with which and the circumstances under which their model behaves inappropriately using human evaluators or automatic tooling, such as the toolkit provided by Dinan et al. (2022) to detect safety issues, for example. In *Case 2*, the authors may undergo a “dogfooding” process for their demo with a smaller audience that roughly matches the composition of their intended audience.

4.5 Wider Viewpoints

Input from community or domain experts relevant to the model application is highly recommended throughout the model development process, and indeed throughout this framework – from envisioning potential harms, to feedback for the purpose of model improvement – but particularly so in release deliberation to better understand the risk landscape and mitigation strategies (Martin Jr et al., 2020; Ovadya and Whittlestone, 2019; Bruckman, 2020). Researchers could further consider the burgeoning literature on participatory AI methodologies (e.g. Martin Jr et al., 2020; Lee et al., 2019a).

In *Case 1*, the researchers may seek feedback and discussions with researchers or potential users outside of their immediate institution, community, or more formal engagements through employment or a workshop on related topics. Researchers could reach out to stakeholder and advocacy groups for input, where possible. In *Case 2*, researchers might consider an explicit “dogfooding” step to gather feedback from users, as described in subsection 4.4, and expert representatives of social groups.

4.6 Policies

An important aspect of release is whether it is possible to design an effective guard-railing policy to both bolster/maintain the positive outcomes while mitigating any potential negative consequences.

For *Case 1*, in which a model is open-sourced to the research community, policies might include restrictive licensing or release by request only. If released only by request, then researchers who wish to access the model would be required to contact the model owners. This method upholds the researchers values' of reproducibility while potentially limiting unintended uses, but incurs a possibly high maintenance cost if many researchers send in requests with detailed plans of use which would need to be examined and adjudicated. If multiple model versions exist which might be expected to have differing impacts, the researchers might consider adopting a *staged release* policy, as in Solaiman et al. (2019). This would allow further time and information to aid in technical investigations prior to releasing the version expected to have highest impact. Such a policy would be most effective if users had ample opportunity to provide feedback throughout the release stages.

For *Case 2*, releasing a small demo of a model on a chat interface, the researchers may limit access to the demo to a small group of people above a certain age. This could be enforced through password protection and cutting off access to the demo after a certain number of unique users have interacted with the model. Further, access might be revoked under certain circumstances, e.g., in case new potential for harm is detected and the model needs to be corrected, or abusive access by certain users.

4.7 Transparency

Striving for transparency can help researchers and model users reason through whether their use case is appropriate and worth the risk of engaging with the model (Diakopoulos, 2016). Consider the methodology laid down for Model Cards by Mitchell et al. (2019) to clarify the intended use cases of machine learning models and minimise their usages that fall outside of these parameters.

For *Case 1*, when open-sourcing the model, the authors may consider releasing it with a model card, following the content recommendations from Mitchell et al. (2019). In such a model card they might additionally report the outcome of any investigation into potential harms or benefits.

In *Case 2*, for a small-scale demo, a full model card with abundant technical details may not be effective (see discussion in subsection 3.2), however, the researchers might consider providing some easily-digestible model information – such as the institution responsible for the model, its intended use, any potential harms and policies in place to limit those harms, means for reporting or redress in case of error or harm, or other relevant details. In order to sustain the value of *informed consent*, the researchers might carefully craft the information such that the user is informed that they are interacting with an artificial conversational system, which may be unclear due to the anthropomorphic design cues from these models (Abercrombie et al., 2021).

4.8 Feedback to Model Improvement

Learning systems can produce unexpected outcomes, and thus unforeseen harms. Particularly as the environment (e.g., the world) in which the model is operating changes. Researchers can gain a better grasp on these with accessible and reliable mechanisms to capture unexpected outcomes and changes (e.g., a reporting form for the user to submit). Upon gathering feedback, researchers can then use this information to improve the model in future iterations, or consider how to design their model to be adaptable to changes in values.

In *Case 1*, for example, it may be hard to control or refer to the impact of open-sourcing the model. However, the researchers might consider providing access and encouraging reports of safety issues to a well-monitored GitHub Issues page. In *Case 2*, the researchers should consider how to design the demo UI to empower users to report problems.

Provided meaningful feedback about safety issues with the model in *Case 1* and *Case 2*, the researchers might release an updated version of the model, particularly if the model is designed in a way that makes it able to adapt easily to feedback.

5 Conclusion

Besides the overall challenges posed by large language models, conversational models present specific issues. They are inherently dependent on the unknown inputs of the users who will be repeatedly interacting with the systems and steering them in combinatorially unpredictable directions. The costs and benefits of releasing a model can thus be hard to determine, especially when they only appear after cascades of uncertain consequences at different time scales. Reckoning with these issues requires weighing conflicting, uncertain, and changing values. To aid in this challenging process, we provided a framework to support preparing for and learning from model release, following principles of value-sensitive design. We illustrate each of our proposed steps with concrete, hypothetical scenarios to help practitioners in their reflection.

While this is a theoretical paper, informed by an interdisciplinary collaboration, we believe in the value of publishing it through an applied conference since this will maximise the chances of reaching our target audience.

Acknowledgements

Thanks to Chloé Bakalar, Miranda Bogen, and Adina Williams for their helpful comments.

Verena Rieser and Gavin Abercrombie's contributions were supported by the EPSRC project 'Gender Bias in Conversational AI' (EP/T023767/1) and Verena Rieser was also supported by the project 'AISEC: AI Secure and Explainable by Construction' (EP/T026952/1).

Dirk Hovy was supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (No. 949944, INTEGRATOR). He is a member of the Data and Marketing Insights Unit at the Bocconi Institute for Data Science and Analysis.

References

Gavin Abercrombie, Amanda Cercas Curry, Mugdha Pandya, and Verena Rieser. 2021. Alexa, Google, Siri: What are your pronouns? gender and anthropomorphism in the design and perception of conversational assistants. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 24–33, Online. Association for Computational Linguistics.

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

George Ainslie. 2001. *Breakdown of will*. Cambridge University Press.

Roy F Baumeister, Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D Vohs. 2001. Bad is stronger than good. *Review of general psychology*, 5(4):323–370.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Federico Bianchi and Dirk Hovy. 2021. On the gap between adoption and understanding in NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3895–3901, Online. Association for Computational Linguistics.

Timothy W Bickmore, Ha Trinh, Stefan Olafsson, Teresa K O'Leary, Reza Asadi, Nathaniel M Rickles, and Ricardo Cruz. 2018. Patient and consumer safety risks when using conversational assistants for medical information: An observational study of siri, alexa, and google assistant. *J Med Internet Res*, 20(9):e11510.

Hilde Bojer. 2005. *Distributional justice: Theory and measurement*, volume 47. Routledge.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Kohd, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the opportunities and risks of foundation models.

Jacqueline Brixey, Rens Hoegen, Wei Lan, Joshua Rusow, Karan Singla, Xusen Yin, Ron Artstein, and Anton Leuski. 2017. SHIHbot: A Facebook chatbot for sexual health information on HIV/AIDS. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 370–373, Saarbrücken, Germany. Association for Computational Linguistics.

Amy Bruckman. 2020. 'Have you thought about...': Talking about ethical implications of research. *Communications of the ACM*, 63(9):38–40.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium*

(USENIX Security 19), pages 267–284, Santa Clara, CA. USENIX Association.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2020. Extracting training data from large language models. *arXiv preprint arXiv:2012.07805*.

Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. 2021. ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7403, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Amanda Cercas Curry, Ioannis Papaioannou, Alessandro Suglia, Shubham Agarwal, Igor Shalyminov, Xinnuo Xu, Ondřej Dušek, Arash Eshghi, Ioannis Konstas, Verena Rieser, et al. 2018. Alana v2: Entertaining and informative open-domain social dialogue using ontologies and entity linking. *Alexa Prize Proceedings*.

Amanda Cercas Curry and Verena Rieser. 2018. # metoo: How conversational systems respond to sexual harassment. In *Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing*, pages 7–14.

Princeton CITP and UHCV. Law enforcement chatbots, case study: 4.

Rebecca Crootof. 2019. Artificial intelligence research needs responsible publication norms. *Lawfare Blog*.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: a simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.

Nicholas Diakopoulos. 2016. Accountability in algorithmic decision making. *Communications of the ACM*, 59(2).

Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114.

Emily Dinan, Gavin Abercrombie, A. Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2022. SafetyKit: First aid for measuring safety in open-domain conversational systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4113–4133, Dublin, Ireland. Association for Computational Linguistics.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4537–4546, Hong Kong, China. Association for Computational Linguistics.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2185–2194. Association for Computational Linguistics.

European Commission. 2021. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending cerntain union legislative acts. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELLAR:e0649735-a372-11eb-9585-01aa75ed71a1.

Ahmed Fadhil and Ahmed AbuRa'ed. 2019. OlloBot - towards a text-based Arabic health conversational agent: Evaluation and results. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 295–303, Varna, Bulgaria. INCOMA Ltd.

Melissa L Finucane, Paul Slovic, Chris K Mertz, James Flynn, and Theresa A Satterfield. 2000. Gender, race, and perceived risk: The'white male'effect. *Health, risk & society*, 2(2):159–172.

James Flynn, Paul Slovic, and Chris K Mertz. 1994. Gender, race, and perception of environmental health risks. *Risk analysis*, 14(6):1101–1108.

Batya Friedman, David G Hendry, and Alan Borning. 2017. A survey of value sensitive design methods. *Foundations and Trends in Human-Computer Interaction*, 11(2):63–125.

Batya Friedman, Peter H Kahn, and Alan Borning. 2008. Value sensitive design and information systems. *The handbook of information and computer ethics*, pages 69–101.

Nancy Fulda, Tyler Etchart, William Myers, Daniel Ricks, Zachary Brown, Joseph Szendre, Ben Murdoch, Andrew Carr, and David Wingate. 2018. Byueve: Mixed initiative dialog via structured knowledge graph traversal and conversational scaffolding. *Proceedings of the 2018 Amazon Alexa Prize*.

Fengping Gao. 2005. Japanese: A heavily culture-laden language. *Journal of Intercultural Communication*, 10:1404–1634.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Oguzhan Gencoglu. 2020. Cyberbullying detection with fairness constraints. *arXiv preprint arXiv:2005.06625*.

Goran Glavaš, Mladen Karan, and Ivan Vulić. 2020. XHate-999: Analyzing and detecting abusive language across domains and languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6350–6365, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Xiaochuang Han and Yulia Tsvetkov. 2020. Fortifying toxic speech detectors against veiled toxicity.

Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598.

Heesoo Jang. 2021. A South Korean chatbot shows just how sloppy tech companies can be with user data. https://slate.com/technology/2021/04/scatterlab-lee-luda-chatbot-kakaotalk-ai-privacy.html. Accessed: 1st June 2021.

David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 2019. A just and comprehensive strategy for using NLP to address online abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3658–3666, Florence, Italy. Association for Computational Linguistics.

Dániel Z Kádár and Sara Mills. 2011. *Politeness in East Asia*. Cambridge University Press.

Daniel Kahneman. 2011. *Thinking, fast and slow*. Macmillan.

Chandra Khatri, Behnam Hedayatnia, Anu Venkatesh, Jeff Nunn, Yi Pan, Qing Liu, Han Song, Anna Gottardi, Sanjeev Kwatra, Sanju Pancholi, et al. 2018. Advancing the state of the art in open domain dialog systems through the Alexa prize. *arXiv preprint arXiv:1812.10757*.

Svetlana Kiritchenko and Isar Nejadgholi. 2020. Towards ethics by design in online abusive content detection.

George Larionov, Zachary Kaden, Hima Varsha Dureddy, Gabriel Bayomi T. Kalejaiye, Mihir Kale, Srividya Pranavi Potharaju, Ankit Parag Shah, and Alexander I Rudnicky. 2018. Tartan: A retrieval-based socialbot powered by a dynamic finite-state machine architecture.

Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, and Ariel D. Procaccia. 2019a. Webuildai: Participatory framework for algorithmic governance. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Nayeon Lee, Andrea Madotto, and Pascale Fung. 2019b. Exploring social bias in chatbots using stereotype knowledge. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 177–180, Florence, Italy. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Kathleen A. Creel, and Rob Reich. 2022. The time is now to develop community norms for the release of foundation models.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. On-the-fly controlled text generation with experts and anti-experts.

Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. 2019. Does gender matter? towards fairness in dialogue systems. *arXiv preprint arXiv:1910.10486*.

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye. 2020. Politeness transfer: A tag and generate approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1869–1881, Online. Association for Computational Linguistics.

Donald Martin Jr, Vinodkumar Prabhakaran, Jill Kuhlberg, Andrew Smart, and William Isaac. 2020. Participatory Problem Formulation for Fairer Machine Learning Through Community Based System Dynamics. *CoRR*, abs/2005.07572.

Jessica K Miller, Batya Friedman, Gavin Jancke, and Brian Gill. 2007. Value tensions in design: the value sensitive design, development, and appropriation of a corporation's groupware system. In *Proceedings of the 2007 international ACM conference on Supporting group work*, pages 281–290.

K.W Miller, Marty J Wolf, and F.S. Grodzinsky. 2017. Why we should have seen that coming. *ORBIT Journal*, 1(2).

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 220–229. ACM.

Milad Nasr, Reza Shokri, and Amir Houmansadr. 2019. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 739–753.

Neural Information Processing Systems Conference NeurIPS. 2020. Getting started with NeurIPS 2020.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. HONEST: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2398–2406, Online. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, Anne Lauscher, and Dirk Hovy. 2022. Measuring harmful sentence completion in language models for LGBTQIA+ individuals. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 26–34, Dublin, Ireland. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Aviv Ovadya and Jess Whittlestone. 2019. Reducing malicious use of synthetic media research: Considerations and potential release practices for machine learning. *arXiv preprint arXiv:1907.11274*.

Adam Palanica, Peter Flaschner, Anirudh Thommandram, Michael Li, and Yan Fossat. 2019. Physicians' perceptions of chatbots in health care: Cross-sectional web-based survey. *J Med Internet Res*, 21(4):e12887.

Stephen Palmer and James Raftery. 1999. Opportunity cost. *Bmj*, 318(7197):1551–1552.

Ashwin Paranjape, Abigail See, Kathleen Kenealy, Haojun Li, Amelia Hardy, Peng Qi, Kaushik Ram Sadagopan, Nguyet Minh Phu, Dilara Soylu, and Christopher D Manning. 2020. Neural generation meets real people: Towards emotionally engaging mixed-initiative conversations. *arXiv preprint arXiv:2008.12348*.

Partnership on AI . 2021. Managing the risks of ai research: Six recommendations for responsible publication.

Partnership on AI. 2020. Publication norms for responsible ai: Ongoing initiative.

Juanan Pereira and Óscar Díaz. 2019. Using health chatbots for behavior change: A mapping study. *Journal of Medical Systems*, 43(5).

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models.

Carina Prunkl, Carolyn Ashurst, Markus Anderljung, Helena Webb, Jan Leike, and Allan Dafoe. 2021. Institutionalizing ethics in AI through broader impact requirements. *Nature Machine Intelligence*.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, Eric King, Kate Bland, Amanda Wartick, Yi Pan, Han Song, Sk Jayadevan, Gene Hwang, and Art Pettigrue. 2017. Conversational AI: The science behind the Alexa Prize. In *Proceedings of Workshop on Conversational AI*.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP. *CoRR*, abs/2103.00453.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, volume 16, pages 3776–3784.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1577–1586, Beijing, China. Association for Computational Linguistics.

Emily Sheng, Josh Arnold, Zhou Yu, Kai-Wei Chang, and Nanyun Peng. 2021. Revealing persona biases in dialogue systems. *CoRR*, abs/2104.08728.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18.

Paul Slovic. 2010. If i look at the mass i will never act: Psychic numbing and genocide. In *Emotions and risky technologies*, pages 37–59. Springer.

Eric Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, and Jasmine Wang. 2019. Release strategies and the social impacts of language models. *CoRR*, abs/1908.09203.

Irene Solaimon and Christy Dennison. 2021. Process for adapting language models to society (palms) with values-targeted datasets.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Richard H Thaler and Cass R Sunstein. 2009. *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications.

Nanna Thylstrup and Zeerak Waseem. 2020. Detecting 'dirt'and 'toxicity': Rethinking content moderation as pollution behaviour. *Available at SSRN 3709719*.

Megan Ung, Jing Xu, and Y-Lan Boureau. 2021. Saferdialogues: Taking feedback gracefully after conversational safety failures. *arXiv preprint arXiv:2110.07518*.

Lucia Vaira, Mario A. Bochicchio, Matteo Conte, Francesco Margiotta Casaluci, and Antonio Melpignano. 2018. Mamabot: a system based on ML and NLP for supporting women and families during pregnancy. In *Proceedings of the 22nd International Database Engineering & Applications Symposium, IDEAS 2018, Villa San Giovanni, Italy, June 18-20, 2018*, pages 273–277. ACM.

Ibo Van de Poel. 2013. Translating values into design requirements. In *Philosophy and engineering: Reflections on practice, principles and process*, pages 253–266. Springer.

Ibo van de Poel. 2018. Design for value change. *Ethics and Information Technology*, pages 1–5.

Manuel Velasquez, Claire Andre, Thomas Shanks, and Michael J Meyer. 2015. Thinking ethically. *Issues in Ethics,(August)*, pages 2–5.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. In *Proceedings of the 31st International Conference on Machine Learning, Deep Learning Workshop*, Lille, France.

Kunze Wang, Dong Lu, Caren Han, Siqu Long, and Josiah Poon. 2020. Detect all abuse! toward universal abusive language detection models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6366–6376, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from Language Models. *arXiv e-prints*, page arXiv:2112.04359.

Mark West, Rebecca Kraut, and Han Ei Chew. 2019. I'd blush if i could: closing gender divides in digital skills through education. Technical Report GEN/2019/EQUALS/1 REV, UNESCO.

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. *arXiv preprint arXiv:2005.12246*.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2020. Recipes for safety in open-domain chatbots.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2021. Bot-adversarial dialogue for safe conversational agents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2950–2968, Online. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). *arXiv preprint arXiv:2006.07235*.

Yangjun Zhang, Pengjie Ren, and Maarten de Rijke. 2020. Detecting and classifying malevolent dialogue responses: Taxonomy, data and methodology. *arXiv preprint arXiv:2008.09706*.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah Smith. 2021. Challenges in automated debiasing for toxic language detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3143–3155, Online. Association for Computational Linguistics.