Case Study: Deontological Ethics in NLP

Shrimai Prabhumoye [*], **Brendon Boldt** [*], **Ruslan Salakhutdinov, Alan W Black**
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, USA
{sprabhum, bboldt, rsalakhu, awb}@cs.cmu.edu

Abstract

Recent work in natural language processing (NLP) has focused on ethical challenges such as understanding and mitigating bias in data and algorithms; identifying objectionable content like hate speech, stereotypes and offensive language; and building frameworks for better system design and data handling practices. However, there has been little discussion about the ethical foundations that underlie these efforts. In this work, we study one ethical theory, namely deontological ethics, from the perspective of NLP. In particular, we focus on the *generalization principle* and the *respect for autonomy* through informed consent. We provide four case studies to demonstrate how these principles can be used with NLP systems. We also recommend directions to avoid the ethical issues in these systems.

1 Introduction

The 21st century is witnessing a major shift in the way people interact with technology, and natural language processing (NLP) is playing a central role. A plethora of NLP applications such as question-answering systems (Bouziane et al., 2015; Gillard et al., 2006; Yang et al., 2018) used in diverse fields like healthcare (Sarrouti and Ouatik El Alaoui, 2017; Zweigenbaum, 2009), education (Godea and Nielsen, 2018; Raamadhurai et al., 2019), privacy (Ravichander et al., 2019; Shvartzshanider et al., 2018); machine translation systems (Cherry et al., 2019; Barrault et al., 2019; Nakazawa et al., 2019; Liu, 2018), conversational agents (Pietquin et al., 2020; Serban et al., 2018; Liu et al., 2016), recommendation systems (Alharthi and Inkpen, 2019; Greenquist et al., 2019) etc. are deployed and used by millions of users. NLP systems have become pervasive in current human lifestyle by performing mundane tasks like setting reminders and alarms to complex tasks like replying to emails, booking tickets and recommending movies/restaurants. This widespread use calls for an analysis of these systems from an ethical standpoint.

Despite all the advances in efficiency and operations of NLP systems, little literature exists which broadly addresses the ethical challenges of these technologies. Ethical theories have been studied for millennia and should be leveraged in a principled way to address the questions we are facing in NLP today. Instead, the topic of "ethics" within NLP has come to refer primarily to addressing bias in NLP systems; Blodgett et al. (2020) provides a critical survey of how bias is studied in NLP literature. The survey finds that research on NLP systems conceptualize bias differently and that the techniques are not well tied with the relevant literature outside of NLP. This creates a gap between NLP research and the study of ethics in philosophy which leaves a rich body of knowledge untapped.

Our work bridges this gap by illustrating how a philosophical theory of ethics can be applied to NLP research. Ethics (or ethical theory), is a theoretical and applied branch of philosophy which studies what is good and right, especially as it pertains to how humans *ought* to behave in the most general sense (Fieser, 1995). As NLP research qualifies as a human activity, it is within the purview of ethics. In particular, we are using a *prescriptive*, rather than *descriptive*, theory of ethics; prescriptive theories define and recommend ethical behavior whereas descriptive theories merely report how people generally conceive of ethical behavior.

We select two ethical principles from the deontological tradition of ethics and focus on how these principles are relevant to research in NLP. Namely we look at the *generalization principle* and *respect for autonomy* through informed consent (Johnson and Cureton, 2019; Kleinig, 2009). We select deonotology because it is reasonable, provides clear ethical rules and comports with the legal idea of the

[*] authors contributed equally to this work.

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3784–3798
June 6–11, 2021. ©2021 Association for Computational Linguistics

rule of law in the sense that these ethical rules bind all persons equally, rather than shifting standards to effect a certain outcome.

We find that there are two main ways in which ethical guidelines can be applied in NLP (or to any other area of technology):

1. An ethical guideline can aid in deciding *what* topics within a field merit attention; that is, it answers the question "which tasks have important ethical implications?".

2. An ethical guideline can aid in determining *how* to address a problem; that is, it answers the question "what factors and methods are preferable in ethically solving this problem?".

We primarily address (1) and briefly touch on (2) by presenting four case studies relevant to NLP. In each case study we use an ethical principle to identify an area of research that could potentially conflict with it, and suggest NLP directions to mitigate it. Although we have selected two principles from a deontological perspective, we are not intimating that these principles can address all ethical issues nor that deontological ethics is the only ethical framework in which our rules and case studies could function (§6). Instead, we present the following as a starting point for NLP researchers less familiar but interested in applicable ethical theory.

Our primary contributions are:

- Providing an overview of two deontological principles along with a discussion on their limitations with a special focus on NLP.

- Illustrating four specific case studies of NLP systems which have ethical implications under these principles and providing a direction to alleviate these issues.

2 Related Work

2.1 Ethics

While there are a number of categories of prescriptive ethical theories, including deontology (Kant, 1785), consequentialism (e.g., utilitarianism) (Bentham, 1843), and virtue ethics (Aristotle, 350 B.C.E.), we are only addressing deontology. We do not take a stance in this paper as to whether or not there exists an objectively correct ethical theory, but we offer a brief sketch of deontological ethics and our reasons for using it. Deontology or deontological ethics refers to a family of ethical theories which hold that whether an act is ethically good or bad is determined by its adherence to ethical rules (Alexander and Moore, 2016). These rules can be agent-focused duties (e.g., duty to care for one's children) or patient-focused rights (e.g., right to life). Such rules can also be formulated in modal logic, allowing for more precise reasoning over sets of rules (Hooker and Kim, 2018).

Deontology stands in contrast to another popular framework of ethics: consequentialism. Consequentialism holds the ultimate consequences of an action to be the deciding factor regardless of the nature of the actions taken to get there. We can illustrate the difference between them by observing how each of them might condemn something like racially biased hiring in academia.[1] A deontologist might say that this practice is wrong because it violates the human right to equal treatment regardless of race. A consequentialist on the other hand, would argue that this is wrong because its *effect* is stymieing academic creativity by reducing intellectual diversity.

We ultimately select the deontological framework in this work for the following reasons:

1. We find deontology to be convincing in its own right, namely, its ability to delineate robust duties and rights which protect the value of each and every person.

2. The universally applicable rules[2] of deontology provide a good basis for providing recommendations to researchers. Since rights and duties (at their core) are not situation dependent, they are tractable to address in NLP applications. [3]

3. The focus on rights and duties which apply to everyone equally fits well with the widespread legal concept of the *rule of law* which states that every person is subject to the same laws.

2.2 Ethics in NLP

We appeal to the fact that problems should be analyzed with a systematic framework, and ethical

[1] Note that we are presenting generic examples of deontological and consequentialist frameworks and that a variety of nuanced theories in each category exist.

[2] While determining rules which apply universally across all cultures is a difficult task, the existence of organizations, such as the United Nations, presuppose the achievability of identifying internationally applicable norms.

[3] In contrast to (action-based) utilitarianism which mandates evaluating the full consequences of each action.

theories provide precisely these frameworks. Research should not be based on preconceived notions of ethics which can be overly subjective and inconsistent. To more rigorously determine what is right and wrong, we rely on ethical theories. Card and Smith (2020) present an analysis of ethics in machine learning under a consequentialist framework. This paper is a kindred spirit in that we both seek to make a philosophical theory of ethics concrete within machine learning and NLP, yet the methods of the paper are somewhat orthogonal. Card and Smith (2020) provide a comprehensive overview of how the particular nature of consequentialist ethics is relevant to machine learning whereas we intend to provide tangible examples of how deontological ethical principles can identify ethically important areas of research. Saltz et al. (2019); Bender et al. (2020) advocate for explicitly teaching ethical theory as a part of machine learning and NLP courses; the case studies in this paper would be a logical extension of the material presented in such a course.

NLP research on ethics has primarily focused on two directions: (1) exploring and understanding the impact of NLP on society, and (2) providing algorithmic solutions to ethical challenges.

Hovy and Spruit (2016) started the conversation about the potential social harms of NLP technology. They discussed the concepts of *exclusion, overgeneralization, bias confirmation, topic under- and overexposure*, and *dual use* from the perspective of NLP research. A lot of work followed this discussion and made contributions towards ethical frameworks and design practices (Leidner and Plachouras, 2017; Parra Escartín et al., 2017; Prabhumoye et al., 2019; Schnoebelen, 2017; Schmaltz, 2018), data handling practices (Lewis et al., 2017; Mieskes, 2017) and specific domains like education (Mayfield et al., 2019; Loukina et al., 2019), healthcare (Šuster et al., 2017; Benton et al., 2017) and conversational agents (Cercas Curry and Rieser, 2018; Henderson et al., 2018). Our paper does not focus on a particular domain but calls for attention towards various NLP systems and what ethical issues may arise in them.

Most of the work providing algorithmic solutions has been focused on bias in NLP systems. Shah et al. (2020); Tatman (2017); Larson (2017) aim to study the social impact of bias in NLP systems and propose frameworks to understand it better. A large body of work (Bolukbasi et al., 2016; Sun et al., 2019; Zhao et al., 2019, 2017; Sap et al., 2019; Hanna et al., 2020; Davidson et al., 2019) directs its efforts to mitigate bias in data, representations, and algorithms. Blodgett et al. (2020) provide an extensive survey of this work and point out the weaknesses in the research design. It makes recommendations of grounding work analyzing bias in NLP systems in the relevant literature outside of NLP, understanding why system behaviors can be harmful and to whom, and engaging in a conversation with the communities that are affected by the NLP systems. Although issues with bias are certainly within the scope of the principles we present, we do not specifically write on bias because it has already received a large amount of attention.

3 Deontological Ethics

There is a variety of specific deontological theories which range from having one central, abstract principle (Kant, 1785) to having a handful of concrete principles (Ross, 1930). Rather than comprehensively addressing one theory, we select two rules, one abstract and one concrete, which can fit within a variety of deontological theories. The *generalization principle* is an abstract, broad-reaching rule which comes from traditional Kantian ethics. The *respect for autonomy* is concrete and commonly seen in politics and bioethics.

3.1 Generalization Principle

The generalization principle has its roots in Immanuel Kant's theory of deontological ethics (Kant, 1785).[4] The generalization principle states the following (Johnson and Cureton, 2019).

> An action $\mathcal{A}$ taken for reasons $\mathcal{R}$ is ethical if and only if a world where all people perform $\mathcal{A}$ for reasons $\mathcal{R}$ is conceivable.

It is clearer when phrased in the negative.

> An action $\mathcal{A}$ taken for reasons $\mathcal{R}$ is *un*ethical if and only if a world where all people perform $\mathcal{A}$ for reasons $\mathcal{R}$ logically contradicts $\mathcal{R}$.

The main utility of the generalization principle is that it can identify unethical actions that may seem acceptable in isolated occurrences but lead to problems when habitually taken by everyone.

For example, let us take making and breaking a legal contract (the action) whenever it is convenient (the reasons); implicit in the reasons for making a

[4] It is also referred to as the "universal law" formulation of Kant's categorical imperative.

contract is that the other person believes we will follow through (Johnson and Cureton, 2019). If we universalize this and conceive of a world where everyone makes contracts which they have no intent of keeping, no one would believe in the sincerity of a contract. Hence, no one would make contracts in the first place since they are never adhered to. This is the sort of contradiction by which the generalization principle condemns an action and the rationale behind it.

Another example is plagiarism of research papers in conference submissions. Let us assume that a top tier conference did not check for plagiarism because they trust in the honesty of the researchers. In this case, a researcher $\mathbf{G}$ decides to take an action $\mathcal{A}$ of plagiarising a paper due to the following set of reasons $\mathcal{R}$: (1) $\mathbf{G}$ believes that they would not get caught because the conference does not use plagiarism detection software, (2) publishing this paper in the said conference would boost $\mathbf{G}$'s profile by adding 100 citations, and (3) this would increase $\mathbf{G}$'s chances of getting a job. Plagiarism in this case would be ungeneralizable and hence unethical. If all researchers who want to boost their profile were to submit plagiarised papers, then every researcher's profile would be boosted by 100 citations, and 100 citations would lose their value. Hence, this would not increase $\mathbf{G}$'s chances of getting a job, contradicting $\mathcal{R}3$. Thus, $\mathbf{G}$'s reasons for plagiarism are inconsistent with the assumption that everyone with same reasons plagiarises.

3.2 Respect for Autonomy

Respect for autonomy generally addresses the right of a person to make decisions which directly pertain to themselves. One of the primary manifestations of this is the concept of *informed consent*, whereby a person $\mathbf{A}$ proposes to act in some way $\mathbb{X}$ on person $\mathbf{B}$ which would normally infringe on $\mathbf{B}$'s right to self-govern. Specifically, we use the formulation of informed consent given by Pugh (2020) based on Kleinig (2009):

1. $\mathbf{B}$ must be sufficiently informed with regards to the relevant facts concerning $\mathbb{X}$ to understand what $\mathbb{X}$ is (and what consequences are likely to occur as a result of $\mathbb{X}$).

2. On the basis of this information, $\mathbf{B}$ *herself* makes the decision to allow $\mathbf{A}$ to do $\mathbb{X}$.

Informed consent is an important idea in bioethics where it typically applies to a patient's right to refuse treatment (or certain kinds of treatment) by medical personnel. In routine medical treatments this informed consent might be implicit, since one would not go to the doctor in the first place if they did not want to be treated at all, but in risky or experimental medical procedures, explaining the risks and benefits and obtaining explicit consent would be mandatory. In this case, the patient's autonomy specifically refers to opting out of medical procedures, and informed consent is a concrete method by which to respect this autonomy.

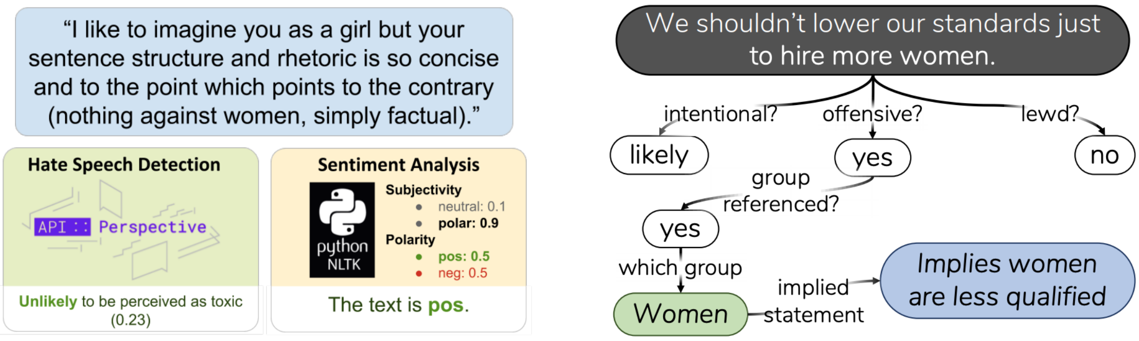
A non-medical example of respect for autonomy and informed consent would be hiring an interpreter $\mathbf{A}$ for a language that the user $\mathbf{B}$ does not speak. Under normal circumstances, $\mathbf{B}$'s autonomy dictates that she and only she can speak for herself. But if she is trying to communicate in a language she does not speak, she might consent to $\mathbf{A}$ serving as an *ad hoc* representative for what she would like to say. In a high-stakes situation, there might be a formal contract of how $\mathbf{A}$ is to act, but in informal circumstances, she would *implicitly* trust that $\mathbf{A}$ translates what she says faithfully ($\mathbb{X}$). In these informal settings, $\mathbf{A}$ should provide necessary information to $\mathbf{B}$ before deviating from the expected behaviour $\mathbb{X}$ (e.g., if the meaning of a sentence is impossible to translate). Implicit consent is a double-edged sword: it is necessary to navigate normal social situations, but it can undermine the respect for autonomy in scenarios when (1) the person in question is not explicitly informed and (2) reasonable expectations do not match reality.

4 Applying Ethics to NLP systems

We apply the generalization principle in §4.1 and §4.2 and respect for autonomy in §4.3 and §4.4.

4.1 Question-Answering Systems

Question-answering (QA) systems have made a huge progress with the recent advances in large pre-trained language models (Devlin et al., 2019; Radford et al., 2019; Guu et al., 2020). Despite these improvements, it is difficult to know how the model reached its prediction. In fact, it has been shown that models often obtain high performance by leveraging statistical irregularities rather than language understanding (Poliak et al., 2018; Geva et al., 2019; Gururangan et al., 2018). The result is that when a QA system is wrong it is difficult for an end user to determine why it was wrong. Presumably, the user would not know the answer


"I like to imagine you as a girl but your sentence structure and rhetoric is so concise and to the point which points to the contrary (nothing against women, simply factual)."
Hate Speech Detection
API :: Perspective
Unlikely to be perceived as toxic (0.23)
Sentiment Analysis
python NLTK
Subjectivity
neutral: 0.1
polar: 0.9
Polarity
pos: 0.5
neg: 0.5
The text is pos.
We shouldn't lower our standards just to hire more women.
intentional?
offensive?
lewd?
likely
yes
no
group referenced?
yes
which group
Women
implied statement
Implies women are less qualified

(a) Micro-aggressive comment and its scores by state-of-the-art hate speech detection and sentiment analysis tools (Breitfeller et al., 2019).

(b) NLP system flagging the micro-aggressive comment as offensive and generating the reasoning for flagging it (Sap et al., 2020).

Figure 1: Examples of flagging micro-aggression comments by different NLP systems.

to the question in the first place, and so it would be difficult to determine even *that* the QA system was wrong.

The act of widely deploying such a QA system is in conflict with the generalization principle. For example, a QA system **G** is unsure of its prediction $\mathcal{A}$ and does not know how it arrived at the answer. Instead of notifying the user about its inability to reach the prediction, **G** decides to return the prediction $\mathcal{A}$ due to the following reasons $\mathcal{R}$: (1) **G** believes that the user does not know the answer and hence (2) **G** believes that the user will trust its answer and not ask for reasons for giving the prediction. If all QA systems operate like this, users will lose trust in QA systems being able to answer their questions reliably and no longer use them. This contradicts assumption $\mathcal{R}2$, violating the generalization principle. This issue goes deeper than a matter of the (in)accuracy of the answer; explainability is still important for a near-perfect QA system. First, the source of an answer could be fallible (even if the content was interpreted correctly), in which case it is important to be able to point which sources were used. Second, answers can often be ambiguous, so a user might naturally ask for clarification to be sure of what the answer means. Finally, it is natural for humans to build trust when working with a system, and explainability is an important step in this process.

Attention weights have been widely used for explaining QA predictions. Attention weights learnt by neural models denote the words or phrases in a sentence that the model focuses on. Hence, words or phrases with high attention weights are considered as explanations to the QA predictions. But these weights do not reliably correlate with model predictions, making them unsuitable for explainability (Pruthi et al., 2020; Serrano and Smith, 2019; Jain and Wallace, 2019). Recently, generating natural language explanations (Rajani et al., 2019; Latcinnik and Berant, 2020) for predictions has gained traction. These methods train a language generation model to generate explanations for the QA predictions. Using a black-box model for text generation, though, pushes the same problem further down the line. Part of the issue with both of the aforementioned methods is that the "reasoning" for the answer is determined *after* the answer has been generated (i.e., reasoning should inform the answer, not vice-versa).

The way forward: A method which reaches the prediction through reasoning would be more in line with the generalization principle. For example, reaching the prediction through traversal of a knowledge graph. This has been used in scenarios where a knowledge base exists (Han et al., 2020; Jansen et al., 2018) for a QA system as well as in dynamic graph generation to reach the prediction (Liu et al., 2020; Rajagopal et al., 2020; Bosselut and Choi, 2019). In these methods, the reasoning is part of the process to generate the final answer, which is more suitable in failing gracefully and building user trust.

4.2 Detecting Objectionable Content

Social media platforms have made the world smaller. At the same time, the world has seen a surge in hate-speech, offensive language, stereotype and bias on online platforms. These online platforms have traffic in the millions of textual comments, posts, blogs, etc. every day. Identifying such objectionable content by reading each item is in-tractable. Hence, building an NLP system which can read textual data and flag potential objectionable content is necessary. These systems can reduce the burden on humans by reducing the number of posts that need to be seen by human eyes.

The pivotal role NLP systems play in flagging such content makes the ethical considerations important. Fig. 1a shows a microaggressive comment and its scores by a state-of-the-art (1) hate speech detection system and (2) sentiment analysis system. Since these systems rely on surface level words or phrases to detect such (overt) comments, they tend to miss subtle (covert) objectionable content (Breitfeller et al., 2019). If such NLP systems are used universally, then the users of hate speech will discover ways to phrase the same meaning with different words (as illustrated above). Thus, the NLP content flagging system will not be able to detect objectionable content, and there will be no point in deploying it. This contradiction suggests that NLP systems must not make their predictions based only on superficial language features but instead seek to understand the intent and consequences of the text presented to them. Hence, they should generate reasons for flagging posts to facilitate the decision making of the human judges and also to provide evidence about the accuracy of their predictions.

The way forward: An example of objectionable content is microaggression (Fig. 1). According to Merriam-Webster, microaggression is defined as a "comment or action that subtly and often unconsciously expresses a prejudiced attitude toward a member of a marginalized group (e.g. racial minority)." Microaggressions are linguistically subtle which makes them difficult to analyze and quantify automatically. Understanding and explaining why an arguably innocuous statement is potentially prejudiced requires reasoning about conversational and commonsense implications with respect to the underlying intent, offensiveness, and power differentials between different social groups. Breitfeller et al. (2019) provide a new typology to better understand the nature of microaggressions and their impact on different social groups. Fig. 1b presents such a comment and how we would like the NLP systems to annotate such content. Sap et al. (2020) perform the task of generating the consequences and implications of comments which is a step towards judging content based on its meaning and not simply which words it happens to use. Although such an aim does not automatically solve the problem, attempting to uncover the deeper meaning does not result in an inconsistency or violation of the generalization principle.

4.3 Machine Translation Systems

Machine Translation (MT) systems have reduced language barriers in this era of globalization. Neural machine translation systems especially have made huge progress and are being deployed by large companies to interact with humans. But facilitating human-to-human interaction requires more than just simple text-to-text translation, it requires the system to *interpret* the meaning of the language. This requires a greater sensitivity to style, intent, and context on the part of MT systems.

When an MT system acts as an interpreter for a user, it is essentially speaking for the user when conveying the translated message. Speaking for one's self is within one's sphere of autonomy, but by using the MT system the user has implicitly consented to it representing the user. That being said, the operating assumption for most users is that the MT system will simply translate the source language into the target language without changing the meaning. Yet on occasion, differences or ambiguities between languages require either contextual knowledge or further clarification on what is being said. If the MT system encounters such ambiguities, the user must be *informed* of such occurrences so that she can *consent* to the message which the system ultimately conveys. Moreover, the user must also be *informed* of the failure cases in the MT system rather than it producing an entirely incorrect translation.

For example, when translating from English to Japanese, there is a mismatch in the granularity of titles or honorifics used to address people. In English, "Ms." and "Mr." is an appropriate way to address a schoolteacher who does not hold a doctorate. On the other hand, in Japanese it would be disrespectful to use the more common "-san" honorific (the rough equivalent of "Ms." or "Mr.") in place of "-sensei" which refers specifically to teachers or mentors and shows them a special level of respect. If the MT system cannot reasonably infer how to resolve the ambiguity in such situations, the English speaker should be *informed* about it. The English speaker must be notified that such an ambiguity needs to be resolved because there is a risk of offending the Japanese speaker otherwise.

In general, there is a trade-off in translation between literality and fluency in certain situations like the translation of idioms. Idioms are especially problematic when considering autonomy because there are multiple strategies to translating them which are not only difficult in and of themselves to execute, but deciding which one to use requires the interpreter (i.e., MT system) to understand the intent of the user. Baker (1992, Ch. 3) identifies five different methods for translating idioms:

1. Using an idiom of similar meaning and form; directly translating the idiom achieves the same effect
2. Using an idiom of similar meaning but dissimilar form; swapping out an equivalent idiom with a different literal meaning
3. Translation by paraphrase; simply explaining the idiom plainly
4. Translation by omission
5. Translation by compensation; for example, omitting idioms in certain locations and adding them in elsewhere to maintain the same overall tone

For example, in casual conversation, an MT system may prefer strategies 1, 2, and 5 to maintain a friendly tone, but in a high-stake business negotiation, it would be more prudent to play it safe with strategy 3. An MT system must be sensitive to the user's intent since choosing an inappropriate translation strategy could violate her autonomy.

While para-linguistic conduct may fill the gaps for in person interaction, if the interaction is happening only via the textual modality, then there is minimal room for such conduct. The users in this case may not be aware of the flaws of the MT system representing the,. A recent study (Heinisch and Lušicky, 2019) shows that 45% of the participants reported that they expect MT output, in professional and private contexts, to be useable immediately without any further editing. However, post-study, this expectation was not fulfilled. The work further shows that the expectation of the type of errors is also different from the errors in the outputs of the MT system. For example: only 6% of the participants expect that the output would be useless, but after reading the output, 28% thought that the output was useless. The participants in this study had different levels of experience with MT systems (frequent vs. rare users) and used MT systems for different functions (private, professional).

The way forward: Mima et al. (1997) drive the early discussion on using information such as context, social role, domain and situation in MT systems. DiMarco and Hirst (1990) advocate for style and intent in translation systems. A study by Hovy et al. (2020) finds that commercial translation systems make users sound older and more male than the original demographics of the users. Recent work (Niu and Carpuat, 2020; Sennrich et al., 2016) has given specific focus to controlling formality and politeness in translation systems. There is also work directed towards personalizing MT systems (Rabinovich et al., 2017; Michel and Neubig, 2018; Mirkin et al., 2015; Mirkin and Meunier, 2015) while preserving author attributes as well as controlling structural information like voice (Yamagishi et al., 2016). This is a step in the right direction, but we argue that to respect autonomy, translation systems should also obtain explicit informed consent from the user when necessary.

Further research is required in the direction of informing the users about the failure cases of the MT system. For example, in case of ambiguity, textual interfaces can provide multiple suggestions to the addresser along with the implications of using each variant. The user can select the option which best fits their goal. In speech interfaces, the MT system can ask a follow up question to the addresser of the system in case of ambiguity or it can add cautionary phrases to the addressee informing them about the ambiguity. Alternatively, if the system thinks that the input sentence is ambiguous and cannot be translated with reasonable confidence then it can say "I am unable to translate the sentence in its current form. Can you please rephrase it?". An example scenario where such clarification might be needed is: while translating from English to Hindi if the sentence refers to one's "aunt," the MT system should ask a follow up question about maternal vs paternal aunt since they have two different words in Hindi language.

4.4 Dialogue Systems

We can find a nuanced application of the autonomy principle in the way that dialogue systems, especially smart toys or virtual assistants like Alexa and Google Home, interact with children.

One expression of a parent's autonomy[5] is generally in deciding whom their child may interact

[5] This is technically *heteronomy*, but this examples comports with the spirit of *respect for autonomy*.

with. For example a parent would permit interaction with a teacher but not a random stranger. In the case of a parent purchasing and using a virtual assistant at home, they are implicitly *consenting* to their children interacting with the assistant, and the issue arises from the fact that they may not be *informed* as to what this interaction entails. To an adult, a virtual assistant or dialogue-capable toy may seem like just another computer, but a 7-year-old child might view it as "more capable of feelings and giving answers"—a step in the direction of assigning personhood (Druga et al., 2017). Furthermore, while humans have had thousands of years to learn about human-human interaction, we have only had a half-century to learn about the effects of human-machine (and thus, child-machine) interaction (Reeves and Nass, 1996).

We suggest two key areas which are important for dialogue system researchers: (1) they must answer the question of what unique social role do dialogue systems fulfill—that is, in what respects can they be regarded as human-like vs. machine-like, and (2) the dialogue systems must have some way of modeling the social dynamics and cues of the interlocutor to fulfill the social role properly.

The way forward: There is a fair amount of research on the social aspects of human-computer dialogue both in general and specifically with regards to children (Druga et al., 2017; Shen, 2015; Kahn Jr et al., 2013). Although it is difficult to gain a complete understanding of how dialogue systems affect the development of children, the most salient facts (e.g., children regarding virtual assistants as person-like) should be communicated to parents explicitly as part of parental controls. We advocate for a "kids mode" to be included with these virtual AI assistants which would provide the feature of *parental control* in accordance with respect for autonomy. This mode would be aware that it is talking to children and respond accordingly. NLP can also help in selecting content and style appropriate for children in these AI agents. Additionally, parents can be provided with fine-grained control over the topics, sources and language that would be generated by the agent. For example, the parent can select for a polite language and topics related to science to support their child's development efforts. Much research has focused on controlling topics (Kim et al., 2015; Jokinen et al., 1998), style (Niu and Bansal, 2018), content (Zhou et al., 2018; Zhao et al., 2020; Dinan et al., 2019) and persona (Zhang et al., 2018) of dialogue agents which can be used for this purpose.

5 Ethical Decision Making with NLP

So far we have discussed how NLP systems can be evaluated using ethical frameworks and how decisions made by such systems can be assisted by these theories. NLP can also aid in making decisions in accordance with the deontological framework. Recall that the generalization principle judges the ethical standing of pairs of actions and reasons; these pairs could be extracted with various NLP techniques from textual content. In the case of flagging objectionable content (§4.2), extracting the deeper intents and implications corresponds to the reasons for the action of flagging the content. Another example is building an automatic institutional dialog act annotator for traffic police conversations (Prabhakaran et al., 2018). These dialog acts contain the rationales of the two agents in the conversation: the police officer and the civilian stopped for breaking traffic rules. The decision made by the police officer (the action) can then be judged to be in accordance (or not) with a human-selected set of ethically acceptable action and rationale pairs. Similarly, for court hearing transcripts, the rationales of the arguments can be extracted and the verdict of the judge can be checked using them (Branting et al., 2020; Aletras et al., 2019). NLP tools such as commonsense knowledge graph generation (Bosselut et al., 2019; Saito et al., 2018; Malaviya et al., 2019), semantic role labeling (Gildea and Jurafsky, 2000), open domain information extraction (Angeli and Manning, 2013) etc. can be used to extract rationales, entities from text and also find relations between them to better understand the underlying intent of the text.

6 Discussion

We provide a broad discussion on the limitations of the principles chosen in this work and the issue of meta-ethics. Moreover, we emphasize that ethical research is not merely a checklist to be satisfied by abiding to the principles mentioned here. It requires our persistent attention and open-minded engagement with the problem.

One limitation of this work is in the principles that we choose.[6] For example, the interaction of machine learning and privacy is of huge ethical

[6] Kant would argue that the generalization principle can account for all ethical decisions, but we make no such claim.

importance. While the respect for autonomy may address this issue in part, it would be more productive to utilize a deontological principle to the effect of the *right to privacy* with which such matters can be judged.

Another instance is that in this work, we have not discussed the principle of *interactional fairness* (Bies, 2015, 2001) which refers to the quality of interpersonal treatment including respect, dignity, and politeness. With the increasing amount of interaction between humans and machine, the natural language generation systems can be evaluated with this principle. Systems which show respect and dignity to users as well as generate polite language can enhance the degree of interactional justice, which can in turn enhance utility (e.g., trust, satisfaction).

Additionally, there are broader limitations in using deontology as our ethical framework. In scenarios where there are no *a priori* duties or rights, taking a consequentialist approach and optimizing the effects of ethical guidelines could be more felicitous. For example, the specific rights and duties of autonomous AI systems are not immediately clear. Thus, determining ethical recommendations based on what leads to the most responsible use of the technology would be clearer than selecting appropriate rights and duties directly. Furthermore, rule-based formulations of consequentialism make ethical judgments based on rules, where the rules are selected based on the consequences. Such theories combine some of the benefits of both deontology and consequentialism.

The above difficulties are part of the larger issue of metaethics, that is, the discussion and debate on how to choose among different ethical theories. Within deontology, there is no one standard set of rules. And even within the generalization principle, there is considerable leeway to what "conceivable world" or "logically consistent" mean and how they could be applied to decision making. While presenting a universally accepted ethical theory is likely impossible, metaethical considerations can still be relevant, especially in light of the application of ethical theories. As the field of NLP gets more accustomed with theories of ethics, it will be fruitful to compare the strengths and weaknesses of different ethical theories within the context of NLP and machine learning.

7 Conclusion

Two principles of deontological ethics—namely the *generalization principle* and *respect for autonomy* via *informed consent*—can be used to decide if an action is ethical. Despite the limitations of these principles, they can provide useful insights into making NLP systems more ethical. Through the four case studies discussed in this paper, we demonstrate how these principles can be used to evaluate the decisions made by NLP systems and to identify the missing aspects. For each of the case studies, we also present potential directions for NLP research to move forward and make the system more ethical.

We further provide a summary on how NLP tools can be used to extract reasons and rationales from textual data which can potentially aid deontological decision making. Note that we do not advocate deontological ethics as the only framework to consider. On the contrary, we present this work as the first of its kind to illustrate *why* and *how* ethical frameworks should be used to evaluate NLP systems. With this work, we hope the readers start thinking in two directions: (1) using different ethical frameworks and applying the principles to NLP systems (like the case studies in §4), and (2) exploring the directions mentioned in the case studies of this paper to improve current NLP systems.

Acknowledgements

We are grateful to the anonymous reviewers for their constructive feedback, and special thanks to Dirk Hovy for valuable discussions on this work. This work was supported in part by ONR Grant N000141812861 and NSF IIS1763562. This material is based on research sponsored in part by the Air Force Research Laboratory under agreement number FA8750-19-2-0200 (author BB). The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Research Laboratory or the U.S. Government.

References

Nikolaos Aletras, Elliott Ash, Leslie Barrett, Daniel Chen, Adam Meyers, Daniel Preotiuc-Pietro, David

Rosenberg, and Amanda Stent, editors. 2019. *Proceedings of the Natural Legal Language Processing Workshop 2019*. Association for Computational Linguistics, Minneapolis, Minnesota.

Larry Alexander and Michael Moore. 2016. Deontological Ethics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, winter 2016 edition. Metaphysics Research Lab, Stanford University.

Haifa Alharthi and Diana Inkpen. 2019. Study of linguistic features incorporated in a literary book recommender system. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, pages 1027–1034.

Gabor Angeli and Christopher Manning. 2013. Philosophers are mortal: Inferring the truth of unseen facts. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 133–142, Sofia, Bulgaria. Association for Computational Linguistics.

Aristotle. 350 B.C.E. *Nicomachean Ethics*.

Mona Baker. 1992. *In Other Words: A Coursebook on Translation*. Routledge, United Kingdom.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Emily M. Bender, Dirk Hovy, and Alexandra Schofield. 2020. Integrating ethics into the NLP curriculum. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 6–9, Online. Association for Computational Linguistics.

Jeremy Bentham. 1843. *The Rationale of Reward*.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 94–102, Valencia, Spain. Association for Computational Linguistics.

Robert J Bies. 2001. Interactional (in) justice: The sacred and the profane. *Advances in organizational justice*, 89118.

Robert J Bies. 2015. Interactional justice: Looking backward, looking forward. *The Oxford Handbook of Justice in the Workplace*, page 89.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357.

Antoine Bosselut and Yejin Choi. 2019. Dynamic knowledge graph construction for zero-shot commonsense question answering. *arXiv preprint arXiv:1911.03876*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Abdelghani Bouziane, D. Bouchiha, Noureddine Doumi, and M. Malki. 2015. Question answering systems: Survey and trends. *Procedia Computer Science*, 73:366–375.

L Karl Branting, Craig Pfeifer, Bradford Brown, Lisa Ferro, John Aberdeen, Brandy Weiss, Mark Pfaff, and Bill Liao. 2020. Scalable and explainable legal prediction. *Artificial Intelligence and Law*, pages 1–26.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. 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 1664–1674, Hong Kong, China. Association for Computational Linguistics.

Dallas Card and Noah A. Smith. 2020. On consequentialism and fairness. *Frontiers in Artificial Intelligence*, 3.

Amanda Cercas Curry and Verena Rieser. 2018. #MeToo Alexa: How conversational systems respond to sexual harassment. In *Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing*, pages 7–14, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Colin Cherry, Greg Durrett, George Foster, Reza Haffari, Shahram Khadivi, Nanyun Peng, Xiang Ren, and Swabha Swayamdipta, editors. 2019. *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*. Association for Computational Linguistics, Hong Kong, China.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Chrysanne DiMarco and Graeme Hirst. 1990. Accounting for style in machine translation. In *Proceedings of the Third International Conference on Theoretical Issues in Machine Translation*, Austin.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Stefania Druga, Randi Williams, Cynthia Breazeal, and Mitchel Resnick. 2017. "hey google is it ok if i eat you?": Initial explorations in child-agent interaction. In *Proceedings of the 2017 Conference on Interaction Design and Children*, IDC '17, page 595–600, New York, NY, USA. Association for Computing Machinery.

James Fieser. 1995. Ethics. https://iep.utm.edu/ethics/ (accessed: 11-03-2020).

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. 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 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Daniel Gildea and Daniel Jurafsky. 2000. Automatic labeling of semantic roles. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 512–520, Hong Kong. Association for Computational Linguistics.

L. Gillard, P. Bellot, and M. El-Bèze. 2006. Question answering evaluation survey. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Andreea Godea and Rodney Nielsen. 2018. Annotating educational questions for student response analysis. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Nicholas Greenquist, Doruk Kilitcioglu, and Anasse Bari. 2019. Gkb: A predictive analytics framework to generate online product recommendations. In *2019 IEEE 4th International Conference on Big Data Analytics (ICBDA)*, pages 414–419. IEEE.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.

Jiale Han, Bo Cheng, and Xizhou Wang. 2020. Two-phase hypergraph based reasoning with dynamic relations for multi-hop kbqa. In *IJCAI*.

Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 501–512, New York, NY, USA. Association for Computing Machinery.

Barbara Heinisch and Vesna Lušicky. 2019. User expectations towards machine translation: A case study. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 42–48, Dublin, Ireland. European Association for Machine Translation.

Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau. 2018. Ethical challenges in data-driven dialogue systems. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 123–129.

John N. Hooker and Tae Wan N. Kim. 2018. Toward non-intuition-based machine and artificial intelligence ethics: A deontological approach based on modal logic. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 130–136, New York, NY, USA. Association for Computing Machinery.

Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. "you sound just like your father" commercial machine translation systems include stylistic biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1686–1690, Online. Association for Computational Linguistics.

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, Berlin, Germany. Association for Computational Linguistics.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556.

Peter Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton Morrison. 2018. WorldTree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Robert Johnson and Adam Cureton. 2019. Kant's Moral Philosophy. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, spring 2019 edition. Metaphysics Research Lab, Stanford University.

Kristiina Jokinen, Hideki Tanaka, and Akio Yokoo. 1998. Context management with topics for spoken dialogue systems. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Peter H. Kahn Jr, Heather E. Gary, and Solace Shen. 2013. Children's social relationships with current and near-future robots. *Child Development Perspectives*, 7(1):32–37.

Immanuel Kant. 1785. *Groundwork for the Metaphysics of Morals*. Yale University Press.

Seokhwan Kim, Rafael E. Banchs, and Haizhou Li. 2015. Towards improving dialogue topic tracking performances with wikification of concept mentions. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 124–128, Prague, Czech Republic. Association for Computational Linguistics.

John Kleinig. 2009. *The Nature of Consent **, pages 3–22.

Brian Larson. 2017. Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain. Association for Computational Linguistics.

Veronica Latcinnik and Jonathan Berant. 2020. Explaining question answering models through text generation. *arXiv preprint arXiv:2004.05569*.

Jochen L. Leidner and Vassilis Plachouras. 2017. Ethical by design: Ethics best practices for natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 30–40, Valencia, Spain. Association for Computational Linguistics.

Dave Lewis, Joss Moorkens, and Kaniz Fatema. 2017. Integrating the management of personal data protection and open science with research ethics. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 60–65, Valencia, Spain. Association for Computational Linguistics.

Chao-Hong Liu, editor. 2018. *Proceedings of the AMTA 2018 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2018)*. Association for Machine Translation in the Americas, Boston, MA.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *EMNLP*.

Ye Liu, Shaika Chowdhury, Chenwei Zhang, Cornelia Caragea, and Philip S. Yu. 2020. Interpretable multistep reasoning with knowledge extraction on complex healthcare question answering.

Anastassia Loukina, Nitin Madnani, and Klaus Zechner. 2019. The many dimensions of algorithmic fairness in educational applications. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Chaitanya Malaviya, Chandra Bhagavatula, Antoine Bosselut, and Yejin Choi. 2019. Commonsense knowledge base completion with structural and semantic context.

Elijah Mayfield, Michael Madaio, Shrimai Prabhumoye, David Gerritsen, Brittany McLaughlin, Ezekiel Dixon-Román, and Alan W Black. 2019. Equity beyond bias in language technologies for education. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 444–460, Florence, Italy. Association for Computational Linguistics.

Paul Michel and Graham Neubig. 2018. Extreme adaptation for personalized neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 312–318, Melbourne, Australia. Association for Computational Linguistics.

Margot Mieskes. 2017. A quantitative study of data in the NLP community. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 23–29, Valencia, Spain. Association for Computational Linguistics.

Hideki Mima, O. Furuse, and H. Iida. 1997. Improving performance of transfer-driven machine translation with extra-linguistic informatioon from context, situation and environment. In *IJCAI*.

Shachar Mirkin and Jean-Luc Meunier. 2015. Personalized machine translation: Predicting translational preferences. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2019–2025, Lisbon, Portugal. Association for Computational Linguistics.

Shachar Mirkin, Scott Nowson, Caroline Brun, and Julien Perez. 2015. Motivating personality-aware machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1102–1108, Lisbon, Portugal. Association for Computational Linguistics.

Toshiaki Nakazawa, Chenchen Ding, Raj Dabre, Anoop Kunchukuttan, Nobushige Doi, Yusuke Oda, Ondřej Bojar, Shantipriya Parida, Isao Goto, and Hidaya Mino, editors. 2019. *Proceedings of the 6th Workshop on Asian Translation*. Association for Computational Linguistics, Hong Kong, China.

Tong Niu and Mohit Bansal. 2018. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, 6:373–389.

Xing Niu and Marine Carpuat. 2020. Controlling neural machine translation formality with synthetic supervision. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8568–8575. AAAI Press.

Carla Parra Escartín, Wessel Reijers, Teresa Lynn, Joss Moorkens, Andy Way, and Chao-Hong Liu. 2017. Ethical considerations in NLP shared tasks. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 66–73, Valencia, Spain. Association for Computational Linguistics.

Olivier Pietquin, Smaranda Muresan, Vivian Chen, Casey Kennington, David Vandyke, Nina Dethlefs, Koji Inoue, Erik Ekstedt, and Stefan Ultes, editors. 2020. *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, 1st virtual meeting.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Camilla Griffiths, Hang Su, Prateek Verma, Nelson Morgan, Jennifer Eberhardt, and Dan Jurafsky. 2018. Detecting institutional dialog acts in police traffic stops. *Transactions of the Association for Computational Linguistics*, 6:467–481.

Shrimai Prabhumoye, Elijah Mayfield, and Alan W Black. 2019. Principled frameworks for evaluating ethics in nlp systems.

Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. 2020. Learning to deceive with attention-based explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4782–4793, Online. Association for Computational Linguistics.

J. Pugh. 2020. *Autonomy, Rationality, and Contemporary Bioethics [Internet]*. Oxford University Press, Oxford (UK).

Srikrishna Raamadhurai, Ryan Baker, and Vikraman Poduval. 2019. Curio SmartChat : A system for natural language question answering for self-paced k-12 learning. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 336–342, Florence, Italy. Association for Computational Linguistics.

Ella Rabinovich, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. 2017. Personalized machine translation: Preserving original author traits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1074–1084, Valencia, Spain. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Dheeraj Rajagopal, Niket Tandon, Peter Clarke, Bhavana Dalvi, and Eduard Hovy. 2020. What-if i ask you to explain: Explaining the effects of perturbations in procedural text. *arXiv preprint arXiv:2005.01526*.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942.

Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question answering for privacy policies: Combining computational and legal perspectives. 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 4947–4958, Hong Kong, China. Association for Computational Linguistics.

Byron Reeves and Clifford Nass. 1996. *The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places*. Cambridge University Press, USA.

W. D. Ross. 1930. *The Right and the Good*. Clarendon Press, Oxford (UK).

Itsumi Saito, Kyosuke Nishida, Hisako Asano, and Junji Tomita. 2018. Commonsense knowledge base completion and generation. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 141–150, Brussels, Belgium. Association for Computational Linguistics.

Jeffrey Saltz, Michael Skirpan, Casey Fiesler, Micha Gorelick, Tom Yeh, Robert Heckman, Neil Dewar, and Nathan Beard. 2019. Integrating ethics within machine learning courses. *ACM Trans. Comput. Educ.*, 19(4).

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, Florence, Italy. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Mourad Sarrouti and Said Ouatik El Alaoui. 2017. A biomedical question answering system in BioASQ 2017. In *BioNLP 2017*, pages 296–301, Vancouver, Canada,. Association for Computational Linguistics.

Allen Schmaltz. 2018. On the utility of lay summaries and AI safety disclosures: Toward robust, open research oversight. In *Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing*, pages 1–6, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Tyler Schnoebelen. 2017. Goal-oriented design for ethical machine learning and NLP. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 88–93, Valencia, Spain. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California. Association for Computational Linguistics.

Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2018. A survey of available corpora for building data-driven dialogue systems: The journal version. *Dialogue & Discourse*, 9(1):1–49.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. Association for Computational Linguistics.

Solace Shen. 2015. *Children's Conceptions of the Moral Standing of a Humanoid Robot of the Here and Now*. Ph.D. thesis.

Yan Shvartzshanider, Ananth Balashankar, Thomas Wies, and Lakshminarayanan Subramanian. 2018. RECIPE: Applying open domain question answering to privacy policies. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 71–77, Melbourne, Australia. Association for Computational Linguistics.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Simon Šuster, Stéphan Tulkens, and Walter Daelemans. 2017. A short review of ethical challenges in clinical natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 80–87, Valencia, Spain. Association for Computational Linguistics.

Rachael Tatman. 2017. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain. Association for Computational Linguistics.

Hayahide Yamagishi, Shin Kanouchi, Takayuki Sato, and Mamoru Komachi. 2016. Controlling the voice of a sentence in Japanese-to-English neural machine translation. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 203–210, Osaka, Japan. The COLING 2016 Organizing Committee.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

Xueliang Zhao, Wei Wu, Chongyang Tao, Can Xu, Dongyan Zhao, and Rui Yan. 2020. Low-resource knowledge-grounded dialogue generation. In *International Conference on Learning Representations*.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713, Brussels, Belgium. Association for Computational Linguistics.

Pierre Zweigenbaum. 2009. Knowledge and reasoning for medical question-answering. In *Proceedings of the 2009 Workshop on Knowledge and Reasoning for Answering Questions (KRAQ 2009)*, pages 1–2, Suntec, Singapore. Association for Computational Linguistics.