A Comprehensive Review of Ethical Frameworks in Natural Language Processing

Peter Francis North

A research Paper submitted to the University of Dublin, in partial fulfilment of the requirements for the degree of Master of Science Interactive Digital Media

2019

Declaration

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at:
http://www.tcd.ie/calendar

I have also completed the Online Tutorial on avoiding plagiarism 'Ready, Steady, Write', located at http://tcd-ie.libguides.com/plagiarism/ready-steady-write

I declare that the work described in this research Paper is, except where otherwise stated, entirely my own work and has not been submitted as an exercise for a degree at this or any other university.

Signed: _____

Peter Francis North
10 May 2019

Permission to lend and/or copy

I agree that Trinity College Library may lend or copy this research Paper upon request.

Signed: _____

Peter Francis North

10 May 2019

Acknowledgements

I would like to thank my supervisor, Dr. Carl Vogel, for all the guidance he provided to support me in my task of writing this dissertation. I would also like to thank my parents, my brother, and my sister, who were always only a phone call away when I needed some advice. Finally, I would like to thank my wife, Marianne, for her unwavering belief in me during this undertaking.

Abstract

An abundance of variation can be applied to the method of experimentation and problem solving that is carried out in the field of Natural Language Processing (NLP). This can be seen through an evaluation of the current literature. An intrinsic characteristic to any method of experimentation, within the field of NLP and without, is to adhere to a morally and legally appropriate framework embedded in ethics. Depending on the approaches seen in the literature, it is often possible to determine individual characteristics within those frameworks that adhere to the mandates of ethical principles. By evaluating the ethical considerations as seen within the literature of NLP, and providing an overview of ethical theories, I attempt to categorise the literature into ethical frameworks whose methods correspond to those theories. I conclude that the vast majority of literatures appear to employ Consequentialist frameworks that are primarily concerned with providing an ideal environment in which NLP can progress.

Keywords: Natural Language Processing, Ethics, Ethical Theories, Ethical Frameworks, Ethical Considerations

Contents

1 Introduction

Natural Language Processing (NLP) is an area of Computer Science concerned with the interaction between computers and natural languages so that natural languages can be analysed and manipulated in order to perform various tasks (Chowdhury, 2003). There are many different subtopics associated with NLP, including, but not limited to, Natural Language Understanding (NLU), Natural Language Generation (NLG), Part-of-speech tagging, coreference resolution, sentiment analysis and speech recognition. Each of these subtopics can be subdivided further into categories, depending on the linguistic area to which they apply.

In order to fulfil their duty in advancing the state-of-the-art, NLP researchers and their experiments often involve an end user. Experimentation can have an impact on that user in terms of specific actions and/or outcomes, some of which can be seen as inappropriate or amoral. As a result of this perception, it is important for NLP practitioners to question how ethical their proposed practices actually are, before carrying out their task (Leidner and Plachouras, 2017). It follows, then, that research is often carried out in accordance to principles that are grounded in ethical theories.

In this review, I will discuss various ethical theories and principles, as well as provide a comprehensive analysis of NLP literature in order to determine if characteristics of those principles can be observed. In section 2, a review of a number of ethical theories is fleshed out. This section includes theories rooted in consequentialism, deontologism, and more modern theories including Computer Ethics.

In section 3, I will attempt to provide an overview of the ethical considerations that are common in relation to Natural Language Processing. These issues are discussed with respect to publications seen within recent years, and from the perspectives of researchers and end-users alike.

In section 4, I will attempt to categorise these publications into ethical frameworks. This categorisation will be based on ethical principles that can be deduced from the approaches to solving NLP issues in the respective literature.

Section 5 will involve a case study on speculation of alternate ethical frameworks with respect to an NLP specific task. I will take this question and attempt to provide a summary of ethical considerations when viewed through various frameworks.

Finally, I will attempt to provide a conclusion of where the vast majority

of the literature stands, in terms of ethical theories that were applied. An overview of future steps to continue the prioritisation of ethical considerations will conclude my discussion.

2 Ethical Frameworks

In order to ensure a sufficient understanding of ethical principles and the ethical issues that arise in the field of NLP, it is important to first provide a definition of the ethical frameworks that are employed. There is notable diversity in the range of approaches to ethics, however within this diversity it is possible to define a number of fundamental similarities. As a result, similar approaches to ethics are often categorised together under an ethical theory, and can broadly be seen as entities belonging to that particular theory. A high-level understanding of these theories is useful in the act of determining which particular ethical frameworks might be employed in NLP tasks, however it is important to be aware of the underlying ethical approaches to truly comprehend the entirety of the framework that is in use. I will now attempt to provide a definition for these theories, before delving more deeply into the ethical approaches associated with each of them.

2.1 Consequentialism

Consequentialist theories can be defined as those that are predominantly concerned with the ethical repercussions of certain actions or activities. Simply put, consequentialism is a doctrine of morality which states that the correct act in any given situation is the one that produces the greatest possible outcome for that situation and for whichever parties are involved (Scheffler, 1988).

One example of consequentialism is Utilitarianism. In keeping with the fundamental nature of consequentialism, utilitarianism has the view that one should aim to maximise the amount of good that can result from a situation, for the good of themselves and the good of others (Driver, 2014). In many respects, utilitarianism seems to be more broadly considered in situations involving a large number of people, where the greatest result for that group is the desired outcome.

Take, for example, a common scenario in NLP: that of shared tasks. Shared tasks are of particular importance as they help to boost the pace of development in the field, which in turn facilitates a culture whereby improving upon

the current state-of-the-art technology is universally the primary concern. (Escartín et al., 2017). In an ideal utilitarian setting, the existence of a medium by which researchers could share tasks confidently while eliminating ethical considerations such as secretiveness or funding bias would exist so that everyone in the community could benefit equally.

A variant of this approach is ethical egoism. This approach is different to utilitarianism in how it perceives the outcome of a situation. While utilitarianism strives to achieve the greatest good for the largest number (Van Staveren, 2007), the egoistic approach modifies the doctrine so that the ultimate goal is to achieve an outcome that maximises self-interest and places little value in the ethical consequences that it might have for others (Rachels, 2012).

This approach can be seen in NLP scenarios where data privacy regulations such as GDPR and HIPAA are seen as obstacles to the progression of Clinical NLP applications. Despite the need for these regulations to protect the individual and their right to privacy, there is the possibility that some may view it merely as a barrier to obtaining clinical language data for training that would otherwise be openly available. An opinion, therefore, that "Perhaps, it is time to rethink the right to privacy in health in the light of recent work in ethics of big data..." might come across as an employment of an egoistic ethical framework, due to apparent disregard of the ethical consequences (Šuster et al., 2017).

2.2 Non-consequentialism

Non-consequentialist theories are fundamentally different to consequentialist theories. Non-consequentialism is less concerned with the ethical consequences of particular actions; instead it focuses primarily on the intentions of the person performing those actions, or the intentions behind their ethical decision making. Unlike consequentialist theories, the emphasis on performing actions in order to obtain the best outcome is lessened significantly, in that there is often little or no value placed on the consequences of an action, rather the value is placed on the action itself (Suzumura and Xu, 2001).

One such example could include a deontological approach to ethics, most often associated with the influential German Philosopher Immanuel Kant (Alexander and Moore, 2016). Also referred to as duty-based ethics, this philosophy ensures that the primary concern lies with the intention behind an ethical decision; in other words, is the decision being made a moral decision? According to Kant's Categorical Imperative, an act may be defined as moral if the act itself

is universally acceptable by everyone, regardless of the consequences (Plaisance, 2007).

An example of this framework might be seen in NLP when an action results in an unforeseen consequence, even if the decision behind the actions which led to that consequence are seen as morally justifiable. Take, for example, a discussion based on the legality of training certain Machine Learning models on data from a selection of annotated corpora. Due to the approach of certain models, it may be considered unethical to have them train on data from certain corpora without firstly obtaining the licencing legally required. If a model is deemed acceptable and acquires licencing to allow for training on certain data, then allowing it to train can be seen as ethically appropriate. Even in the event where the training results in a situation where a re-evaluation of the legalities is eventually required, the decision to allow the training to commence can be seen as ethical, as the framework applied it is primarily concerned with an action that is morally or legally acceptable, and less concerned with any repercussions which might appear (de Castilho et al., 2018).

2.3 Agent-centred Theories

Agent-centred theories differ again when compared with consequentialist and non-consequentialist theories. Whereas consequentialism is primarily concerned with the ethical consequences of particular actions, and non-consequentialism focuses more on the person ethical decisions of the person performing those actions, agent-centred theories are more concerned with the ethical status of individuals, rather than ethical decision making.

An example of this is virtue ethics. Virtue ethics can broadly be identified as one that places value on a person's moral character, an evaluation of how virtuous the character is, as opposed to the consequences of their decisions (Hursthouse, 1999). For those who adhere to a framework bound by virtue ethics, importance is stressed on the character of those individuals or groups, and how it can be improved.

Many thinkers on morality might be inclined to agree that a character's reliability is more significant than their actions in accordance to rules within a defined framework. This reliability in character includes maintaining a good sense of what can be deemed as morally acceptable, as well as responding to ethical situations in an emotionally appropriate manner (Beauchamp and Childress, 2013). When an agent performs some tasks, or responds to a situation in

a way that is deemed morally and socially appropriate, and when that agent is acting in accordance with their own ideals as well, then it may be fair to state that this agent is acting in accordance to virtue ethics. If, however, that agent feels obligated to perform a task, but the action does not reflect their own ideals, then that agent may not be seen to be taking the virtuous approach. Despite their actions being seen as morally acceptable, the fact that they did so against their own inclination can be seen as a negative reflection on their character, and not at all virtuous or reliable (Beauchamp and Childress, 2013). A point about virtue ethics, as discussed, is that the moral nature and development of the agent performing a particular action is what's seen as most valuable (Floridi, 1999).

This can be observed in NLP. An assumption can sometimes be made that virtue ethics applies to a single agent, but due to the subjectivity of an agent, an assumption can also be made that an agent can be defined as a single person or a large group of persons (Floridi, 1999). This includes researchers and applications in the field of NLP, Engineers, or any contributing or non-contributing members of the community. One such example could be the seen in the approach to the developments of modern Artificial Intelligence systems. As time progresses, more and more AI systems are being developed, and this can be seen in the abundance of NLP applications visible in society today. As machine learning allows for AI to learn by experience, for example in pattern recognition, it seems logical that a model could be designed so that an AI could learn from a series of defined moral virtues. As it nears autonomy, this means, in theory, that the AI's character would evolve as it learns more about the moral virtues against which it is being trained. One theory proposes that if a machine is being trained towards attaining a state of temperance, it would lose its "desire for excess of any kind, not even for exponential self-improvement" (Berberich and Diepold, 2018). The theory follows that building virtuous traits into the AI's character would help to avoid losing control of the same AI, and help to achieve a character whose intentions and ideals coincide with virtue ethics.

2.4 Computer Ethics and Information Ethics

In relation to NLP, it might be prudent to discuss also the sphere of Computer Ethics. In its relation to technology and various other fields, unpredictable or unexpected ethical issues are bound to arise. As technology progresses, computers continue to provide us with ever expanding possibilities for action. This

progression, in turn, raises a number of ethical considerations about which policies ought to be constructed and applied with respect to these new possibilities. This has been referred to as a Policy Vacuum (Moor, 1985), where there follows a period of time where rules for how to approach these new possibilities simply do not exist. Some examples of Policy Vacuums created by computers and technology could include the monitoring of employees with potentially intrusive software by employers, or the act of collecting and selling specific data relating to a person and their online browsing habits and transactions, both of which raise ethical and moral questions (Johnson, 2004).

In perhaps a more in-depth discussion on Computer Ethics, Luciano Floridi presents a case of difficulty in its exact classification, due to the fact that many well-versed philosophical theories such as consequentialism and deontologism may not always be considered to be the best option when it comes to dealing with Computer Ethics problems (Floridi, 1999). As such, Floridi proposes that there is a need for a more metaphysical foundation of Computer Ethics, an argument in which harming information itself can be seen as morally questionable (Johnson, 2004). Floridi proposes a theory of Information Ethics, a theory in which the question of what is good for informational entities and the infosphere in general is asked (Floridi, 1999).

According to his argument, Computer Ethics appears to never be primarily concerned with the exact moral and ethical actions in a situation, but rather it concerns itself with what is the "best course of action" after careful analysis, and what is ultimately better for the Infosphere as a whole. Information Ethics can be seen as the philosophical foundation for this analysis, wherein every single entity contained within the infosphere can be seen as an information entity which must be valued; that is to say, every piece of information from the past, present and future, whether it exists or existed in a physical form or not, can be considered an important entity within its own infosphere, and must be treated with respect and in an unbiased fashion (Floridi, 1999). To better describe this concept, Floridi goes into great detail about the intricacies, which are ultimately put to the test against other ethical theories with real world examples.

One such example is that of vandalism. In his example, we consider a boy playing alone in an abandoned dumping-ground, a place in which nobody ever comes. The boy is here for entertainment purposes, and so creates entertainment by smashing windows of abandoned cars with rocks. Although generally considered a negative action in any case, Floridi analyses the boy's vandalism with many ethical theories in mind. Floridi concludes that a theory like conse-

quentialism is difficult when applied to this scenario, as the boy is ultimately having fun and thus generating more happiness in the world, while at the same time doing no harm to anyone else. With virtue ethics in mind, the boy's actions can be seen as a detriment to his overall moral character, as his judgement of the situation and his ethical decisions are seen on a grander social scale as unacceptable and not at all appropriate. In the case of Information Ethics, Floridi argues that it is much easier to define why the boy's vandalism might be condemned: "he is not respecting the objects for what they are", and in entertaining himself in this way he is needlessly disrespecting the infosphere (Floridi, 1999).

A framework of Information Ethics may also be seen within the world of NLP. One consideration within this field is the language data that is openly available. In many scenarios, an analysis of the collection, publishing and availability of research data in publications within the NLP community suggests that many of those publications rely on data that was scraped from the web, but full disclosure on how that information was actually obtained or treated is not provided (Mieskes, 2017). Similarly, repositories linking to datasets that are shared in publication often seem to disappear after a time and is therefore no longer available to the community (Mieskes, 2017). A proposal to create a guide to maximise the protection of this data and the manner in which it is treated could be seen to follow an ethical framework guided by information ethics. This could cover a number of vital questions in order to treat the language data in a respectful and unbiased fashion, such as disclosing how the data was collected, whether or not it has been anonymised if it contained potentially sensitive information, whether or not resulting data has been or will be published, and whether it is possible and will remain possible to obtain the data that was worked with. In this sense, we can then ensure that the infosphere is protected as a whole. By respecting the data and where it came from, and by respecting those who might be interested in findings resulting from those data, all entities involved can be treated in a fair and unbiased capacity.

3 Ethical Considerations

3.1 Overview of Ethical Issues in Natural Language Processing

Like any field of research, NLP has seen its share of ethical issues. These issues come in many different forms. In many cases, some ethical issues appear only after the fact, perhaps as a result some study or work in NLP, and whether or not the issues where anticipated. Other times, researchers actually keep in mind the plethora of ethical issues that have appeared in past works, their intention being to formulate a new and better policy or ethical framework in order to avoid encountering the same issues again. Before we can properly classify the frameworks within which many of these works appear to be bound, it is important to first provide a review of the common ethical considerations and issues that have appeared in NLP. The following section will discuss some of these issues.

It is difficult to categorise the ethical issues I have reviewed on a case by case basis. The majority of issues I have discovered appear to be largely related to how data is used, considerations such as data privacy and anonymisation, the quality of data and the potential for that data to introduce bias into a dataset, and considerations regarding data one must take into account when applying that data for automated tasks through the use of techniques such as machine learning. Rather than categorise the issues into separate sections, I will provide my analysis by going through papers which shared those ethical considerations. The presence of one ethical issue does not automatically discount the existence of another, as issues of fairness can arise from the quality of linguistic data, bias can result from how models are trained, data quality can influence the operability of automation tasks, and many variants in between.

3.2 Ethical issues discovered in the literature

I start my discussion with a focus on issues of data privacy in respect to shared tasks within NLP, such as data transparency, anonymisation and the risk of incurring bias in the dataset. This leads into a discussion on the issues that exist in the procurement of personal data for Clinical purposes. Bias is discussed in great depth from the perspective of various NLP subtopics including speech processing, Natural Language Generation and inference. Applying various Machine Learning techniques to data which is not of sufficient quality is

also discussed, and this includes the difficulties encountered when models are applied across foreign language data. Finally, the many ways in which data is obtained and treated are worth mentioning. Examples include elicitation of language data, lack of clarification of data sources, and failure to apply licensing wherever necessary.

One of the most discussed issues today in the literature is seemingly due to the regulations set forth to protect a person's data, a tremendously important concern in its own right. As a whole, the community goes to great lengths to accommodate fellow researchers and promote competition between each other by sharing language data openly. As a result of the legal situation in Europe at present, however, with the recent introduction of the General Data Protection Regulations, constrictions on copyright protected and/or privacy protected data is stricter than ever. While one might consider these constraints to be essential to the protection of an individual's privacy, they can often times be seen as a barrier to the process of shared task in the NLP community.

One such paper, written by Escartin et. al (Escartín et al., 2017) provides an overview of ethical considerations to ensure an even and fair playing field for shared tasks in NLP. As discussed briefly in a section on ethical theories, shared tasks are almost fundamental for driving NLP research forward. Shared tasks allow researchers to compete with one another in order to receive acknowledgement for their achievements in their respective fields (Escartín et al., 2017). This medium allows researchers to submit systems in order to address specific challenges, and it is commonplace to do so in the field of NLP, particularly in the last number of years at various Computational Linguistics and NLP conferences.

While shared tasks can ultimately be seen as a positive way to encourage contributions and increase competitiveness to move NLP specific issues forward (Escartín et al., 2017), a number of ethical issues are raised in this paper which suggest the current framework counters this, and is somewhat discouraging for certain researchers, depending on their own circumstances. Due to the competitive nature of shared tasks, unethical behaviour in NLP research could be common. An example of this is the act of secretive behaviour regarding one's discoveries or research results, directly defying the encouraged fairness of the shared tasks medium by "unwilling to publish results in a timely fashion, refusal to provide access to data sets and conflicts concerning the 'ownership' of experimental materials" (Escartín et al., 2017). This is a potentially unethical act carried out by the researcher, which actively detracts from the intended effectiveness of shared tasks. By refusing to contribute to the community, hoarding data

prevents progress from being made to that research by like-minded researchers, and ultimately benefits nobody. Similar issues discussed in this paper as a result of the competitive nature of shared tasks also include overlooking the significance of certain ethical considerations, or being influenced either positively or negatively in behaviour as a result of relationships with other researchers (Escartín et al., 2017).

Similar ethical issues can be observed in declining to provide a description of negative findings. Many researchers neglect to see the value in negative results due to fear of diminishing their standing as a scientist. This is often not the case, as there may actually be inherent value for the field in revealing those negative findings. Failure to report such potentially valuable findings might actually be harmful to NLP, as it removes a variable that might otherwise benefit the progression of related research (Escartín et al., 2017).

Such reporting bias can be harmful in various NLP areas. Take Clinical NLP, for example. A simple form of bias might be in the subjectivity of written Natural Language by a trained clinician when taking notes on a patient. There may be significant differences between the words spoken by the patient, and in the way the clinician interprets those words. Similarly, a serious illness may in fact be missed by the clinician, simply due to a negative stigma associated with certain symptoms. A patient's failure to report embarrassing symptoms, therefore, might actually have a negative impact on them in the long run (Šuster et al., 2017).

Similar issues of data privacy and bias can be observed in the article by Tulkens et. al (Šuster et al., 2017). Due to the sensitivity of data today, it is often difficult for Clinical NLP researchers to obtain data for their research within time constraints. As discussed in the previously mentioned article (Escartín et al., 2017), this could be problematic for NLP research. Without consent to obtain the data they require, it is often difficult for researchers to progess and meet previously defined deadlines. Even if consent is eventually obtained, a significant amout of time might have already been spent. In situations where data has been appropriately sanitized, there is a possibility that ethical repercussions could arise from its use. One reason this needs to be considered is due to evidence that suggests sanitized data can often be re-identified later on, despite having been anonymized (Šuster et al., 2017; Hintze and LaFever, 2017). Consent may even still be required even after anonymization. According to the General Data Protection Regulations (GDPR), explicit consent may need to be provided before that data can be officially used (Hintze and LaFever, 2017).

What this means for researchers is that extra time has to be expended in order to gain explicit consent. Without first obtaining this consent, using this data will then be seen as legally and morally inappropriate.

Mieskes et. al (Mieskes, 2017) present a quantitative analysis of publications in the NLP community on the collection, publishing and availability of research data. Evidence may actually suggest that many published works rely on data that has been scraped directly from the internet. Often, this is not disclosed to the reader, and, furthermore, there is a lack of indication as to how that data was actually treated. In many cases, data repositories that were made use of in experiments or other articles seem to disappear after the date of publication, meaning members of the community lose access to that data and may be unable to verify the validity of that data themselves (Mieskes, 2017).

The outcome of this review culminates with a proposal for how to improve this situation within the NLP domain as well as other research areas. To understand this fully, however, it is important to provide an overview of the actual ethical considerations that were discussed. The author makes a point that data is often collected for research without first obtaining consent (Mieskes, 2017). For this reason, they cite the act of collection as a major ethical issue for NLP. It gave rise to considerations of how the data was actually obtained, whether or not the data had been anonymized, whether the results from said data were published, and whether or not it remains possible to gain access to the data.

Anonymization of personal data is of particular importance, as it's misuse could have moral as well as legal implications. As well as revealing information that could be destructive to an individual and their reputation, it could also place an NLP researcher in an awkward position where justification for said actions could be difficult. This is stressed by Mieskes et. al (Mieskes, 2017) as another important ethical consideration for NLP researchers. In the event of publication, a researcher must consider if any sensitive data was used. If so, they must ensure that that data was properly sanitized in order to comply with privacy regulations that are in place.

Data anonymisation has also been criticised for damaging the viability of data so that it may no longer be useful to the researcher. Bias can be introduced to the dataset due to notable differences between those who are willing to participate, and those who are not, i.e. those who agree to provide explicit consent and those who refuse (Šuster et al., 2017). Furthermore, through the exactitude of the natural language used in the explanation of an experiment for which explicit consent is explicitly required, some members of the general

public might refuse to provide consent due to any number of reasons, ranging from misunderstanding of the description to being turned off by mandatory unambiguity of what will be required and/or where their data will be used (Šuster et al., 2017).

Another area of ethical consideration within this paper relates to the social impact of biases in data, where "discrimination can occur when individuals or groups receive unfair treatment as a result of automated processing, which might be a result of biases in the data that were used to train models" (Šuster et al., 2017). One factor is data quality. Unfair biad could be seen when the dataset used in Clinical NLP Inaccurately represents the story being told. If an automated process fails to detect an anomaly in clinical data, it doesn't necessarily mean the process itself is at fault. Rather, it could be due to the quality of data which was provided, in which there was insufficiency of some capacity which meant the model associated with the study could not be properly trained and thus could not sufficiently identify that anomaly (Šuster et al., 2017). Considerations of an ethical, moral and legal nature which can be seen here relate back to the discussion of withholding negative results in shared tasks (Escartín et al., 2017). Further cases of unfairness in automated processes will be discussed in the following paragraphs

Bias can also lead to ethical issues involving gender and dialect. In an experiment by Tatman (Tatman, 2017), an attempt was made to evaluate YouTube's Automated Captioning system across genders and multiple dialects. Results concluded that there may actually be issues which have ethical repercussions for the system, as massive differences in accuracy could be seen depending on gender and dialect. Tatman concludes that a person's sociolinguistic identity may negatively impact their ability to utilise automated speech systems.

This bias has ethical and moral implications for NLP practitioners in their attempt to optimise speech recognition software. It also reflects badly on those from varying sociolinguistic backgrounds, in that they are deprived a service that would, if they were from any other sociolinguistic setting, function to a better standard. It must be noted that results for automated captioning across both genders seemed to be broadly positive, due to a test of pitch wherein the fundamental frequency of a recorded female voice was lowered and compared, seemingly resulting in no real difference in accuracy (Tatman, 2017) which hints that speech processing algorithms are actually successful at capturing varying degrees of voice pitch, whether high or low on the fundamental frequency scale.

Ethical issues, in fact, are more apparent in this experiment when comparing

the word error rate. The results concluded that YouTube's Automated Captioning system had a greater degree of success in processing words spoken by male participants when compared with females. This was also seen across dialects. When bundled together, men and women from Scotland had a greater word error count that men and women from California, and in both regions, women had a greater word error count when compared with men (Tatman, 2017).

This bias raises a number of ethical considerations. Tatman (Tatman, 2017) makes a great point on exposure in her discussion of the experiment. In linguistic terms, no single dialect is any more viable than another, and a person's ability to recognise dialect effectively comes as a result of exposure to that dialect (Tatman, 2017). It follows, then, that training a model on multiple layers of data from many different sociolinguistic backgrounds might eliminate this bias to a certain degree. An ethical consideration for NLP researchers in developing speech processing algorithms might be to ensure there is an appropriate degree of quality in the data on which models are trained. It might be reasonable to compare this with cases of trained data in other Natural Language areas, such as clinical NLP, where datasets have the possibility of lacking essential information and thus often provide an inaccurate representation (Šuster et al., 2017).

Simply the act of classifying gender comes with ethical considerations. By using classification or topic model approaches to objectively categorise gender data can result in drawing stereotypical findings (Koolen and van Cranenburgh, 2017). This may be all the more important in society today due to evidence that suggests people are more likely to use more emotive language in response to women than to men, when categorised and represented in such an objective fashion (Voigt et al., 2018).

Bias can also incur fairness issues. Madnani et. al (Madnani et al., 2017) discusses automated scoring of written and spoken responses in test scenarios. Ethical behaviour must be taken into account to ensure in such an area to ensure all participants in automated scoring be treated fairly. An example of these ethical considerations is in how tests are being scored and interpreted. Factors such as language proficiency or logical reasoning in a test which is not designed to measure either factor could still negatively influence a person's scoring capabilities, especially if the automated scoring system isn't trained to look past those biases. This has ethical implications for the test taker, as an inability to produce a natural language sentence through written or spoken means due to language proficiency might result in a score much lower than that which they are actually capable of obtaining. This places them at a significant disadvantage

when compared with someone whose language proficiency is greater.

Take, for example, a situation in which a company is looking for an individual to join their Data Science division. The first step in securing a position could be through a set of online assessments, in this case designed to provide an estimate of a candidate's analytic skillset. A person with a lower level of language proficiency could be unfairly graded by the system if it is unfairly biased, even in the event that they possess the analytical skills required. Perhaps another person with a higher level of language proficiency could score better in those online assessments, even if their skillset in analytics is much lower than the first candidate. Such a scenario should be considered ethically when designing automated scoring systems. An attempt should be made by NLP practitioners to remove such bias from manual and automated scoring systems so that any test participants, regardless of race, logical reasoning, reading comprehension and language proficiency are fairly scored based purely on the construct against which they are being tested (Madnani et al., 2017). This, again, reflects my discussion on data quality (Šuster et al., 2017), albeit from a different angle.

Fairness in scoring systems can be viewed from a more dangerous perspective. As mentioned in the previous paragraph, one must be wary of the data being used when making use of machine learning techniques to train models. A similar issue is discussed by Schnoebelen (Schnoebelen, 2017) where relying on classification tasks could have terrible ethical repercussions if employed. This comes in the idea of predicting criminality. If we train a model to detect criminality through facial recognition, by training a model on classification based on facial features such as lip curvature, it is inevitable that an innocent group of people will be incorrectly labelled as criminals no matter how effectively the model has been trained (Schnoebelen, 2017). This degree of error can also be seen when applied to NLP. Automated processes can be trained to detect patterns in linguistic data. Common ethical considerations for sentiment analyses and similar tasks could overlook an error which could result in someone being classified as ineligible to apply for a loan, or even racial stereotyping and linguistic profiling (Schnoebelen, 2017).

In many cases, biased labelling or character description in Natural Languages can negatively influence public opinion on individuals or groups of people. Fokkens (Fokkens et al., 2018) focuses on stereotyping of Muslims through something called a microportrait extraction, a method of exploring how a specific group of people are described in Natural Language terms on a wider social scale.

In journalism, one can assume that overgeneralisation of a particular group of people can result in negative repercussions. In NLP, bias can often be seen in words in the form of labelling, or in descriptive characteristics and behaviour when extracted at a sentence level (Fokkens et al., 2018). In terms of labels, this can be seen through stereotypical inferences deduced from nouns, as well as through generic labelling such as "Muslims are..." (Fokkens et al., 2018). In describing characteristics and behaviour, this can often be seen when a distinct characteristic type is associated with a people as a whole, and thus implied as their general behaviour as well. As argued by fokkens et. al (Fokkens et al., 2018), all of the above forms of bias can be negatively influencing, and thus have the potential to create ethical problems for the people who are being stereotyped.

One article by Lohr et. al (Lohr et al., 2018) considers the constrictions of Intellectual Property rights of media content with regard to medical health and clinical language data and the effects they have on the NLP community. Due to legal issues with regard to real world clinical language data in Germany, most importantly the data privacy considerations, Lohr et. al (Lohr et al., 2018) concluded that making use of authentic language data was no longer the most efficient option. As a result, their objective was to counter this restriction by making use of "synthetic" language data instead of real-world data. The optimal result would mean that NLP researchers would no longer be hindered by European regulations on language data, as those privacy concerns would not apply to language data that was artificially manufactured (Lohr et al., 2018).

One needs to consider the ethical repercussions which could arise with the use of real-world language data. As a person essentially owns his or her own data, that means each individual has full control over how any of that data is disclosed, to whomever the data is disclosed, and for what purposes (Fischer-Hübner et al., 2013). For this reason, access to language data in Europe is much more difficult to secure, which poses problems for NLP researchers. Despite the ethical implications of using a person's data without their consent, including legal repercussions for the NLP practitioner and possible harms on the data owner both, a barrier is thus constructed which means that there is a substantial lack of openly available language data for research, particularly for Clinical NLP (Lohr et al., 2018).

This paper focuses on such issues with respect to Non-English language data, German in particular (Lohr et al., 2018). NLP tasks on Non-English language data can often be problematic. For the most part, work in NLP psychotherapy has largely been carried out on English language data, and very little has been

done with German (Mieskes and Stiegelmayr, 2018). This results in a number of ethical considerations for NLP researchers when working with data that is not English. Part-of-speech tagging on syntactic features is often predictive, but they also reveal some problems when working with foreign languages as foreign words can often be mis-labelled as mis-spellings and vice-versa. For reasons such as this, NLP researchers must ensure that transcription is of high enough quality to allow for an efficient Machine Learning based classification technique to be applied (Mieskes and Stiegelmayr, 2018).

Due to the nature of the language style in clinical data, aside from potential issues with language compatibility, existing NLP tools may fall short of being able to properly function. This is because of potential deviation in typography, semantics, syntax and ambiguity in written language data, to name a few (Lohr et al., 2018). An observation has been made that in reusing NLP tools that have been trained prior and primarily on general-language data sources such as newspapers and other data openly available to the public, these models could be seen as lacking in their interpretation of clinical language data due to the vast difference in the language (Hellrich et al., 2015). This again presents a problem for NLP practitioners, as the availability of trained models and various NLP tools may be entirely inappropriate when made use of with radically different language styles.

Another form of ethical consideration in the field of NLP is the way in which language data is obtained for corpora building purposes in Natural Language Generation (Manishina et al., 2016). In order to collect enough data to create or extend corpora in data-driven Natural Language Generation tools, researchers typically make use of two solutions. One solution is to rely on manual paraphrasing of language data, a time consuming and expensive process typically performed by a number of experts. An alternative solution to this is to make use of crowdsourcing, which may raise ethical questions of bias in the collected dataset, or even result in faulty entries that become unusable for the corpus. It is difficult to say which method is more advantageous, considering that crowdsourced data also requires manual validation at a later stage, again a time-consuming process (Manishina et al., 2016).

In order to combat the above issues, a proposal is made to automate the process of corpus extension by integrating automatically obtained synonyms and paraphrases (Manishina et al., 2016). This method would eliminate the need for wasted time and resources by employing experts to manually extend corpora, as well as eliminating potentially unreliable data obtained through

crowdsourcing. It is essential that both methods be considered from an ethical perspective before that extension begins. An expensive option may require a significant portion of funding be expended, and there is no guarantee that crowdsourced data will be more or less reliable.

Braunger et. al (Braunger et al., 2016) observes similar ethical challenges when making use of crowdsourced material for spoken dialog systems in vehicles. While the case can indeed be made in support of using such data, one must always be aware of the exact way in which that data is acquired in order to avoid instances of bias or unusable faulty entries. In comparison to relying manual entry, crowdsourcing can indeed be seen as a cost-effective alternative (Lyding et al., 2018). As pointed out though, this is not always the case. Despite saving on costs and time, crowdsourced data is often frowned upon for producing bad quality data and can add to difficulty as maintaining control over the crowdsourcing environment can be problematic (Braunger et al., 2016).

Spoken dialog systems require a great number of spoken utterances in order to train the speech recognition engine and to determine the overall reliability of a system (Braunger et al., 2016). Braunger et. al (Braunger et al., 2016) present a comparative analysis of three data elicitation methods that are used in crowdsourcing tasks. The objective was to discover which of the three elicitation methods would perform best in the collection of natural language data. The results not only provide a recommendation for which of these methods is best to use, but also clearly highlight potential issues that could appear when using these methods.

To understand the ethical considerations regarding these issues, we first need to briefly examine the elicitation methods which were discussed (Braunger et al., 2016). The difference in the methods lay in how the task was presented to participants, in the form of pictures, semantics, and finally, pure text. In the elicitation task using pictures, participants were given a visual representation of the utterance they should perform. In the semantics representation, participants were instead presented with a number of semantic entities. Finally, in the text representation, participants were given a full natural language sentence defining explicitly what they should do. In each elicitation method, the given tasks were the same (Braunger et al., 2016). In order to determine which of the three elicitation methods would be most appropriate to perform the task of collecting speech utterances, the most frequent utterances provided in the experiment were then analysed, as well as he notable differences that could be seen on both sentence and word level (Braunger et al., 2016).

Based on the analysis, Braunger et. al (Braunger et al., 2016) make a recommendation that the text representation method might be the most appropriate. This recommendation is made on the basis that "it is a good compromise between a high rate of valid data, the linguistic variance and the possibility of creating very specific tasks for different types of commands" when compared with the remaining two elicitation methods (Braunger et al., 2016). In other words, making use of the text representation method might be more reliable in the fact that more speech utterances are considered valid and are not thrown out due to being labelled as faulty, the text representation method allows for more variation in the utterances provided by the participants, in that their capacity to formulate more valid and different sentences for specific tasks is increased. This conclusion might have come about as a result of ethical consideration. By finding a "compromise" (Braunger et al., 2016) between the elicitation methods, an environment securing maximum valid data for researchers as well as a slightly less error prone medium for participants was recommended.

Other ethical considerations can appear for NLP practitioners in terms of legality. For example, the legality of using certain data for experimentation often appears in a grey area, or it may simply be difficult to deduce whether or not that data can actually be used for input in experimentation. This can be taken a step further when considering the legality of using Machine Learning to train models from various resources such as an annotated corpus. This ethical consideration is discussed in a paper by Eckart de Castilho et. al (de Castilho et al., 2018).

We have seen before where making use of machine learning to train a model can lead to ethical repercussions. For example, in the event that models used for speech recognition are not trained on data that is sufficient to recognise dialects across varying sociolinguistic backgrounds, many groups can find themselves at a disadvantage when it comes to making use of that software (Tatman, 2017). Other observable ethical considerations include concerns such as the legalities surrounding the use of pre-trained models, and whether or not authorization is required prior to training models from corpora, whether licensing must be assigned and whether or not the licensing of the corpora affects the licensing applied to the model (de Castilho et al., 2018). Such considerations can mean it may be more ethically appropriate to decide against using a certain model for specific NLP tasks, otherwise there may be a risk of ethical repercussions for the researcher or for any other it might affect.

In summary, ethical considerations can arise from a wide array of variables

within NLP. Ethical issues can appear as a result of ongoing experimentation within the field, but often the need to evaluate the current set of ethical considerations inspires discussions which promote alternative methods and objectives. In the next section, we will re-examine the literature in an attempt to determine the ethical frameworks which match these tasks and outcomes.

4 Ethical Frameworks

4.1 Overview of Ethical Frameworks

In light of the ethical considerations we have discussed, it may be interesting to attempt to categorise the literature with respect to the ethical theories discussed in the previous chapter. By analysing the literature of NLP in relation to ethical theories, one can attempt to determine the ethical framework employed by researchers in their approach to accomplishing their goals. It is possible that the frameworks within which they operate can even influence their findings, and through reflection, a conclusion could also be made that an alternate framework might have been more appropriate. With this in mind, I will attempt to apply various ethical frameworks to the literature I have discussed. In doing so, a greater understanding of the primary motives behind each paper might reveal itself, including how ethical principles may have influenced their approach to completing their experiments and reviews.

4.2 Categorisation of Ethical Frameworks in the literature

Escartin et. al (Escartín et al., 2017) provide a comprehensive review of ethical issues seen in shared tasks for NLP. A proposal to launch a survey follows, within which information about current best practices, recommendations and checklists about issues which should be considered, and whatever information may need to be provided to participants should be included relative to each shared task (Escartín et al., 2017). The ultimate goal of this survey is to improve both shared tasks, as well as the ethical practices employed by those who participate.

A consequentialist ethical framework may fit with this dynamic, as consequentialism can be loosely defined as saying behavioural norms depend only on the consequences that follow, i.e. that the consequence is the only valuable factor (Sinnott-Armstrong, 2015). Many versions of consequentialism can be derived from this, however. What is the result of applying a utilitarian consequentialist framework? Utilitarianism is broadly concerned with producing the

greatest possible outcome that is beneficial to all human participants, the greatest good for the largest number (Van Staveren, 2007). By providing a means to which shared tasks can be considered fair for all researchers, no matter what their circumstances may be, and ideally with the perfect elimination of all ethical problems that might be produced as a result of the competitive nature of shared tasks, one might indeed argue that utilitarianism was the influencing framework in this task. To reinforce this point, a scenario which is fairest for all researchers will also potentially guarantee the best environment for efficient development and progression for NLP research.

With this in mind, however, we may also consider that an ethical framework inspired by Floridi's Information Ethics could be applied (Floridi, 1999). Consider the entire sphere of shared tasks with respect to NLP as the Infosphere. Within Floridi's metaphysical approach to Computer Ethics, one's primary task is to respect the infosphere. This includes any and all informational entities that fall within that infosphere: NLP research, fellow researchers, and even the approach to research practices. By following unethical behavioural patterns such as failing to report negative results and "gaming the system" by focusing exclusively on winning as opposed to obtaining results that are beneficial to NLP as a whole (Escartín et al., 2017), researchers will then be harming the infosphere. By rectifying such problems, by providing a survey of information that may ultimately provide a better medium for shared tasks, the infosphere and its entities can be seen as being respected. This framework, however, does not account for all the issues discussed here. Take, for example, the case of an unequal playing field. If one group of researchers have significantly more processing power at their disposal when compared with a group that is not as ably equipped, the fact that they will outperform their lesser able-equipped fellows does not imply that they have performed unethically. Rather, they have respected the infosphere and ethically made use of the tools that were readily at their disposal.

What of virtue ethics, then? If the actions leading up to the outcome are of equal importance to the outcomes themselves (Koehn, 1995), can we apply an ethical framework close to Aristotelian virtue ethics here? In attempting to improve NLP shared tasks, the authors here demonstrate virtuosity in their approach. By acting in accordance with what is socially acceptable and actively trying to mitigate negative ethical issues as a result of competition, the authors demonstrate reliability of character and thus improve their own character as well. It is easy to see why one might apply this framework here.

These ethical frameworks are not necessarily mutually exclusive, however.

In my opinion, I believe it is fair to argue the case that an ethical framework could be employed that matches the principles behind multiple theories. Utilitarianism and virtue ethics both can be reflected in the approach, as a virtuous outcome is ultimately placed in a position of essential importance.

The ultimate goal of Tulken et. al in their paper on ethical challenges in Clinical NLP (Šuster et al., 2017) was to provide an overview of the ethical challenges that are commonplace in the sphere of Clinical NLP (Šuster et al., 2017). Within this overview, concerns for data privacy are discussed, from a perspective of protecting the individual, as well as appearing as a blocker to NLP progress. Consent and sanitization are reviewed in a similar light. With that in mind, a proposal is set forth to "rethink the right to privacy" (Šuster et al., 2017). How can such an approach be categorised with respect to ethical frameworks?

Perhaps one can consider a Utilitarian ethical framework was employed here. To argue for this point, one has to consider the ultimate goal. In an ideal utilitarian setting, data privacy would be perfectly protected by the need for consent and data sanitization. At the same time, ethical issues for NLP researchers such as wasted time and biased data would be virtually absent. By changing the way in which we approach data privacy, can this ideal scenario be achieved?

Perhaps their approach is idealistically egoistic in its consequentialism. Data privacy regulations such as GDPR and HIPAA are often seen as hurdles which deter the progression of NLP research. Despite the fact that those regulations exist to protect the individual from harm, attempting to promote a change to this policy could be seen to provide an outcome only beneficial to those researchers, perhaps at the risk damaging an individual's right to privacy by removing the need for consent and anonymization.

If we apply Kant's Categorical Imperative, an act can seen as morally sound if the act is considered to be universally acceptable (Plaisance, 2007). From this perspective, perhaps it is the duty of people to provide their data willingly to research, regardless of the outcome. Similarly, one might consider a duty-based framework from the point of view of the researcher, in that acquiring data without consent is only right when it comes to furthering the state-of-the-art. However, this approach doesn't appear to match the framework employed, as Tulken et. al (Šuster et al., 2017) made clear the fact that it is the consequences of these ethical challenges that matter most.

What if we apply Aristotelian virtue ethics? Immediately we can deduce that taking away one's right to consent to their data being used can be seen

as an action that violates traits of virtuosity. In the eyes of the public, this may in fact be viewed as an action which damages both the character of NLP researchers as well as their practices. With this in mind, we can also suppose that Floridi's metaphysical variation on Computer Ethics (Floridi, 1999) was not an influencing factor in their framework, as the suggestion to alter the approach to data privacy could potentially be seen as a disrespectful action against an informational entity.

In their quantitative study on data (Mieskes, 2017), an overview of some of the ethical considerations for NLP practitioners regarding the use of data in light of data privacy regulation was discussed. In order to determine the ethical framework that was most likely employed, we need to apply the ethical theories that were discussed with respect to the method of this article. To conclude this article, a checklist is proposed for authors and reviewers in order to ensure a protective layer covers the data which will be obtained for NLP purposes in future publications (Mieskes, 2017). This checklist includes important ethical considerations such as ensuring data has been collected and used appropriately, and that it will be properly distributed to the community.

With this in mind, one might suppose that a consequentialist framework was employed. It is clear from their approach that the ideal solution is to provide the best possible outcome for individual data owners and NLP researchers alike, where a person's data has been used in accordance with data privacy regulations. Further to this point, one might even go so far as to define this framework as utilitarian. In going out of their way to provide an ethically sound method of data use and publication that protects all parties, and NLP as a whole, Mieskes et. al (Mieskes, 2017) show utilitarian traits by also ensuring a potential way to result in the greatest result for the greatest number (Van Staveren, 2007).

One might also argue in favour of virtue ethics. In recognising the flaws perceived in data usage in NLP research, the act of providing that checklist can be seen to be just as valuable as the desired outcome (Koehn, 1995). It stands to reason that NLP researchers would profit from a safeguard such as the one proposed. By actively seeking to improve the way in which data is obtained, Mieskes et. al (Mieskes, 2017) exhibit virtuous characteristics, and improve their own character by improving the morality of data collection methods for NLP.

We may be able to get a feel for Tatman's (Tatman, 2017) ethical motives behind this experiment by examining the words in her concluding section. She finishes by noting that it would be ideal for automated speech recognition

systems to perform equally in their effectiveness "for users regardless of their sociolinguistic backgrounds", citing also that a deterioration in performance for "disadvantaged groups" only add to the lack of equality (Tatman, 2017). One could easily argue, therefore, that her motives were primarily outcome-based. A consequentialist framework rooted in Utilitarianism seems to fit this approach, as an "ideal" (Tatman, 2017) outcome in this scenario would be mean automated captioning systems would perform better for everyone, regardless of sociolinguistic background, resulting in less marginalisation due to the negation of systemic bias on dialect. The absence of bias would also be beneficial to the NLP engineers behind the algorithms.

Consequentialist egoism fails to match Tatman's approach, as the task is primarily concerned with potentially disadvantaged groups the current algorithms may affect. Perhaps a common good approach can actually be applied, as opposed to utilitarianism, which emphasises respect and compassion for all others, so as to provide conditions which allow for all people within a community to be treated respectfully (Melé, 2009).

Duty-based ethics could be applied to a certain degree, but ultimately falls flat, as an ethical decision might only be morally justifiable if the act itself is acceptable by everyone, regardless of the outcome (Plaisance, 2007). While the action can be seen as important in this scenario, achieving the optimal outcome is valued much greater than the method to which the outcome was achieved.

Madnani et. al (Madnani et al., 2017) suggest that the best way to ensure ethical considerations are made regarding unfair scoring in automated testing, whether from spoken or written responses, is by making automated non-proprietary tools available to NLP researchers that incorporate recommendations and generate analyses which help to resolve issues of bias in scoring systems (Madnani et al., 2017). In terms of ethical frameworks that can be applied to this approach, an emphasis is placed on the humans participating in automated scoring. Value is placed in the consideration that automated scoring systems should be fair, but more emphasis is placed on the fair treatment of the test-takers themselves.

For this reason, one might apply consequentialist frameworks defined by Utilitarian or Common Good principles. A Utilitarian framework might be considered the most ethical of the two, as the common good approach sometimes seems to favour those who are at a disadvantage relative to other parties (Melé, 2009). Utilitarianism, on the other hand, emphasises the ideal outcome for everybody as being the primary goal.

The common good approach might suit works I have discussed such as Koolen et. al (Koolen and van Cranenburgh, 2017) and Voigt et. al (Voigt et al., 2018). In both reports, the notion of bias on gender specific language data is up for question. One might argue that those users who could be negatively affected by objective categorisation of gender by means of sentiment analysis, or in general response terms from other humans via mediums such as reddit, might represent a targeted or disadvantaged group. To ensure this is not the case, Koolen et. al (Koolen and van Cranenburgh, 2017) stress the need for more control over variables than is usually applied within datasets. This act, though, could potentially result in the ideal outcome desired by Utilitarians, as a general decrease in biased interpretation of data across automated categorisation techniques not only applies to those who may appear to be less advantaged.

Similarities are apparent in the discussion on incorporating goal-based design methods into NLP and Machine Learning techniques for the same purposes of ideal outcome (Schnoebelen, 2017).

The paper by Fokkens et. al (Fokkens et al., 2018) can be viewed with many ethical frameworks in mind, based on the various ethical theories discussed in the opening chapter. I will start by trying to apply a framework of consequentialism. Fokkens et. al (Fokkens et al., 2018) stated that the result of the coreference resolution system implemented provides a "new NLP task that is of high interest for researchers". In that sense, this outcome appears to be great for NLP researchers in their attempts to further understand stereotyping language data and their implications. So, what kind of consequentialism might fit best? Utilitarianism is concerned with providing the greatest possible outcome for all people who are involved (Van Staveren, 2007). The outcome in this scenario is directly relevant to NLP researchers who will want to make use of such data. However, this does not mean that it will not be beneficial for the general public at a later stage. If researchers are able to generate a greater understanding of stereotypes taken directly from data that is rife with bias, perhaps someday they could contribute to a system where bias can be reduced, thus reducing the level of negative generalisation.

Perhaps this is more egoistic in its consequentialism, however. It appears that this work has no direct benefit to groups of people who suffer currently due to stereotyping bias. Instead, it provides an alternative approach for researchers in their understanding of stereotyping in Natural Language data. What if we suppose that the coreference resolution system implemented here, while named

"promising" in the article's conclusion (Fokkens et al., 2018), is beneficial to NLP researchers, but is ultimately harmful in reinforcing those stereotypes for groups of people? This, then, would not be an ideal solution for everybody.

Deontologism, in particular something like duty-based ethics, may be difficult to apply, as non-consequentialist frameworks are inherently less concerned with outcomes than the actions themselves (Harris, 2008). The task carried out here appears to be less concerned with the method of determining stereotypes (although they claim the method may be of importance to the field later) as it is with the potential boon the outcome might have for NLP researchers and the social sciences in their attempts to understand stereotypes (Fokkens et al., 2018).

Perhaps, for this reason, one could apply a framework akin to virtue ethics (Koehn, 1995). Value can be seen at multiple facets of this experiment. The actions themselves are of importance as they may provide a means to paint a better picture of what the data is saying in terms of stereotyping. This, again, may actually help in providing a better outcome as well. In this case, both the action and outcome can be seen as ethically and socially appropriate, and reflect well on the researcher's character as well as improving the understanding of the task as a whole.

In our chapter on ethical considerations, I delved into a number of articles that were particularly concerned with the ethical implications that appear as a result of data privacy regulations. In the article written by Lohr et. al (Lohr et al., 2018), three main ethical issues were debated: Intellectual Property Rights, Data Privacy, and complications with applying clinical language data to existing NLP tools. In order to combat these issues, an alternative approach to working with Clinical language data was proposed. To begin with, the suggestion was made to substitute language data in authentic clinical documents with synthetic language data. For example, authentic medical reports with patient-specific language data would be replaced with fake clinical reports which would be published by professionals for educational reasons. In doing so, this would eliminate any privacy concerns of patients, as no real-world language data would be used. Secondly, it is suggested that physical corpus distribution be replaced by sharing software which reliably reconstructs copies of a corpus that is defined in the lab. This, in turn, would help to avoid any Intellectual Property Rights violations as no physical corpus will be distributed (rather a soft copy will be produced using the end-to-end software suite within a laboratory setting) (Lohr et al., 2018). The outcome of this work allowed the researchers

to present the JSYNCC, "the largest and, even more importantly, first publicly available, corpus of German clinical language". This corpus is easily portable to other languages (Lohr et al., 2018).

In order to define the ethical framework employed here, we need to consider who and what is involved. Firstly, we have the researchers themselves. Clinical NLP researchers are hampered by issues such as IPR and data privacy, as well as being unable to make full use of previously used NLP tools due to the different data culture present in clinical data when compared with general-language data (Hellrich et al., 2015). Another important consideration is the patient themselves. By violating IPR and breaching data privacy, the possibility for negative consequences on that patient becomes a possibility. The ultimate goal of the work here seems to be to find a solution that both protects the data holder from harm, while at the same time, providing an effective medium through which Clinical NLP can still progress at a satisfactory pace, thus making the lives of NLP researchers simpler.

One might argue, then, that the ethical approach considered in this work closely resembles consequentialist theories. This is because the researchers themselves seem to be primarily concerned with the ethical consequences of particular actions, consequences such as IPR violations and data privacy breaches. By emphasising goodness of outcomes (Schnoebelen, 2017), they employ an ethical framework closely resembling that of Utilitarianism, with the intention of minimising the potential damage that could occur with respect to IPR violations and data breaches, while providing a way to improve the lives of those whose work has been made more difficult by privacy regulations.

The work carried out by Manishina et. al (Manishina et al., 2016) raises some interesting moral considerations, when considered from an ethical point of view. It is implied that their resulting product can be used for training in other NLG models, as well as use in Natural Language Understanding and other NLP domains. Also discussed is the fact that the corpus extension is much less expensive than manual paraphrasing or crowdsourcing, and does not require human intervention except in the narrowest of circumstances. Ethically speaking then, the development of such a tool could be seen as a boon for researchers and system developers alike. Despite this, however, the tool was created specifically to exclude human input in a system that is built on human language.

One may conclude, however, that the approach here is strongly in line with Utilitarianism. Firstly, the point made about human exclusion as a result of

researchers not using crowdsourced, or experts not being hired to manually append corpora, might be moot. This is because those humans, either hired experts or volunteers in a crowdsourcing exercise, would not be directly involved in the experiment from the beginning. The Utilitarian perspective, however, is not about an action that is carried out. Rather, the consequences which come about as a result of an action are more important. If we look at this issue from Luciano Floridi's metaphysical take on Computer Ethics (Floridi, 1999), those humans could be seen as not having been contained within the defined infosphere. The act of this experiment has no real negative or positive bearing on potential employees or volunteers if they are not sought after in the first place.

From a Consequentialist point of view, the experiment might be seen to fulfil the simplest parameters to match that framework. That is to say, the total outcome of which is obtained through the act of using the automatic tool in this scenario seems to generate an overall state of goodness in relation to the action (Foot, 1985). More specifically, traits of Utilitarianism seem to follow as the results seem to generate the optimal state of affairs for automated corpus extension in NLG, as the results can be seen to achieve the "maximum satisfaction of desire" specific to the scenario (Foot, 1985). This is due to the level of human satisfaction after the fact, as well as the elimination of time-consuming processes and cost.

What kind of ethical framework can be determined by the approach of Braunger et. al (Braunger et al., 2016)? By comparing multiple elicitation methods and recommending one which the experiment suggests yields the best results for crowdsourcing natural language data, the ethical framework employed could be categorised as consequentialist. This point could be argued due to the fact that the potential issues that appear as a result of certain elicitation methods was of primary concern. By making a recommendation on which method appears to provide the best solution for the greatest number (Van Staveren, 2007), in that it is recommended that this approach be used for tasks not only including the collection of speech utterances (Braunger et al., 2016), and even explicitly exclaiming their recommendation as "compromise", one might even go so far as to argue that this is a strong Utilitarian ethical framework (Van Staveren, 2007). In this sense, the outcome can be seen as the fairest outcome for all parties. The text elicitation method is recommended based on the apparent compromise between maximising efficiency in terms of eliciting more valid data and allowing for a greater degree of flexibility in the utterances that are pro-

vided. This is again the best outcome for participants, as it is implied that they feel less limited in their ability to complete each utterance task. This again allows for a greater quality collection of data for researchers to work with in order to further their technology.

It may be unfair to settle on this assumption, as the experiment can be viewed with multiple ethical frameworks in mind. Take virtue ethics, for example. When viewed with Aristotelian virtue ethics, both the outcome and the act can be seen in a positive light, as the action itself is determined to be of the same value as the best result (Koehn, 1995). Both have bearing on the virtuous nature of a particular character, in this case, those who worked towards determining the best solution for crowdsourcing. With this in mind, is it more reasonable to define the ethical framework employed in this experiment as one of virtue ethics?

A number of ethical theories can be applied to the approach laid out by Eckart de Castilho (de Castilho et al., 2018). Eckart de Castilho (de Castilho et al., 2018) provided a legal analysis based on three case studies, in order to determine the safest legal approach for NLP researchers when making use of pre-trained Machine Learning models. By providing this summary, they can perhaps be seen to have used non-consequentialist duty-based ethical principles within their framework. This is due to the emphasis that is placed on the way in which training is executed from a legal perspective. It is less concerned with the ethical repercussions that may occur after the fact, as long as the action of securing authorization and the required licencing has been carried out correctly by NLP practitioners.

5 Case Study: Speculation

In the previous chapter, I attempted to categorise a number of research papers in NLP into ethical frameworks. Variation in this categorisation is apparent, ultimately depending on how individual Natural Language Processing tasks were tackled. One might pose a question as to what would happen if the frameworks were entirely different for each article. What might be considered ethical and unethical if the framework applied was entirely different?

In this short section, I will speculate on the various approaches that could be applied to a specific NLP problem, depending on which framework is applied in order to provide a solution. In the interest of understanding which ethical considerations are at stake, I will apply various ethical frameworks to

that question.

The problem I present is one specific to both Natural Language Generation and Natural Language Understanding: the notion of automated chatbots. Suppose one wanted to introduce chatbots officially in an educational capacity. In order to take this even further for ethical consideration purposes, let us also suppose these chatbots would be introduced in elementary schools and have a direct role in overseeing the education of children. While chatbots have indeed been used in a wide variety of areas including applications such as instant messaging services in order to promote products or elicit services, (Molnár and Szüts, 2018), they have not yet been implemented in schooling systems in such a prominent capacity.

If we view this issue from ethical theories, what questions might we pose in order to see the idea to completion? Ultimately, there are many ethical considerations regarding such a scenario. Would the implementation of such a chatbot result in an outcome where human teachers were no longer required? Would it undermine the schooling system itself, meaning it would be possible to rely on a chatbot for home schooling purposes? For those who utilise such a system for home-schooling, would this detract from their eligibility to apply for a position in university? For the sake of simplicity, I will focus on a narrower set of ethical considerations, those that apply to a system where chatbots are used in the teaching of children in a school setting.

The first step is to summarise the scope of ethical considerations, a number of which will follow here. Assuming such a chatbot could be produced to a sufficient level of quality, is its production ethical? A chatbot of this nature could undermine the role of trained teachers. This could result in ethical issues such as less workable hours, which in turn could lead to lower income. Employability might become even more difficult for teachers.

How about if we view it from the children's perspective? There is always the chance that the chatbot could prove infallible at equipping our children with the skills necessary to progress through the educational system. It follows that the opposite could also be true. One also needs to consider that people learn in many different ways. A teaching method that works for one student may not work for another. It may also be difficult for a child to relate to an Artificial Intelligence when placed in an authoritative role such as this.

What about the chatbot itself? If the chatbot is inadequately trained, a scenario could arise in which the teaching it provides are a detriment to the students it oversees. Perhaps, from a certain point of view, it could give off the

appearance of optimal performance, but at the same time actually deviating from the standard curriculum. The model might also be inherrently biased, resulting in a wider array of ethical issues.

I will place myself in the shoes of an engineer who is tasked with approaching this problem. If I view the issue from a consequentialist perspective, I must focus on the outcome. As a result of this, I must ask questions such as "how can I implement such a chatbot so that I can produce the best overall outcome, in the interest of everyone who is involved in the schooling system(Scheffler, 1988)?". A Utilitarian perspective reinforces this point, as I would need to plan a method which results in the greatest good for the greatest number. This would allow for an outcome with minimal negative repercussions, and maximum value for teachers, students, and the A.I. itself. A Common Good approach would require a framework which emphasises the fairest outcome for all parties involved, particularly when it comes to those who might be perceived as less fortunate, or in some way at a disadvantage. Perhaps, in this sense, greater care would need to be taken to ensure special needs students, or those with issues such as social anxiety, or even autism, would benefit in the same way as those who don't from the lessons provided by the chatbot. This means that the chatbot has to be trained so that it can provide education to all, equally. For the chatbot to fail in this task would be seen as unethical.

What if I employ a non-consequentialist framework? In that case, I might pose a question which places value on steps taken rather than the potential consequences, such as "what techniques will I use when constructing the chatbot?". A deontological framework might view it in such a light, as long as the implementer was motivated by a set of rules that can be seen as morally or socially justifiable (Plaisance, 2007). This can hold true even in the event that the outcome is less than ideal, or even harmful, for the children who are exposed to it. In the event the chatbot performs poorly, and the children learn nothing, as long as their is observable nobility in the intentions behind implementing the chatbot, it can fit within the rules laid out for a duty-based ethical framework, and can therefore still be seen as an ethically justifiable action.

If I was most inspired by Aristotle, and wanted to ensure the framework I followed was defined by virtue ethics, I might combine the two previous questions. This means the action applied should be just as important to me as the outcome: "What are the best techniques I can apply in the implementation of an educational chatbot, so that those techniques result in the best overall outcome for those who will make use of it?". Achieving an idyllic outcome can only

be seen as ethical if the techniques applied in the chatbot's construction are perceived as acceptable. This holds true when reversed, as feeling duty-bound to adhere to a strict approach in it's implementation will always be unethical if the chatbot does not perform to the optimal standard.

Let's consider the model discussed by Floridi (Floridi, 1999). A similar approach would need to be considered here as with a framework bound by Virtue Ethics. The infosphere includes all entities that can be observed in this scenario: the chatbot, students, teachers, the school, the educational system, even the researchers themselves. Will the implementation of a chatbot negatively impact any of these informational entities, in any capacity? If so, then the act will always constitute unethical behaviour. Suppose I implemented a chatbot that performed better than any teacher in any educational setting. If this pushes the teacher out of the picture, I can be seen as performing an act of disrespect. My actions, and the outcomes that follow, will only be considered ethical if every informational entity is respected.

6 Conclusion

In this review, I have attempted to provide a comprehensive review of ethical considerations that appear in the literature of NLP. With ethical theories and principles in mind, as well as the methods, goals and approaches utilised by researchers in each article, I administered a categorisation of these literatures into ethical frameworks. While many of the papers spoke of existing ethical considerations for Natural Language Processing, for the purpose of recommending a framework that could ultimately improve the policies regarding ethics in NLP, many of the outlying papers discussed had goals more specific providing solutions to NLP specific issues. Ethics, while still important in determining their approaches, was secondary to the tasks they set out to solve. In dealing with their own objectives wherein ethics might not have been the primary influencing factor, upon observation once can surmise that they nevertheless acted in accordance with a framework defined by ethical principles.

In section 2, I discussed the various ethical theories that are commonly discussed today, those theories that appear most relevant when it comes to research and production settings for Natural Language Processing. Section 3 underscored a wide variety of ethical considerations and issues that appear in Natural Language Processing, the scope of which led to the categorisation of literature into ethical frameworks in section 4, with respect to those ethical the-

ories that were discussed. Finally, a speculation on the application of alternate ethical frameworks in a case study format was discussed, in order to determine if any difference in the ethical considerations could be perceived.

After providing an extensive review of ethical theories and considerations within NLP, I have come to the conclusion that, at the very least, it is safe to say that none of the articles discussed within had obvious nefarious intentions. Many appear to exhibit qualities commonly associated with Consequentialist theories such as Utilitarianism and Common Good approaches to ethics, as well as similarities to Agent-based theories such as Aristotelian Virtue Ethics. While it is apparent that some of the approaches discussed concerned themselves more with acting in accordance to a duty-bound method, or if the outcome of certain experiments appear to be more egoistically beneficial than ideal for the greatest number of people, I feel confident in summarizing that the general approach to ethics in NLP literature are predominantly guided by morally and socially acceptable principles. This is due to the common theme of betterment across all papers in my dataset, whether that betterment lies in improving existing models to provide better solutions for end-users and researchers alike, or if it furthers the state-of-the-art in NLP in any capacity.

7 Future Work

In order to ensure that updates and improvements continue to be made to Natural Language Processing, it is intrinsic to consistently review and update ethics policies. By reviewing the literature and identifying any ethical issues that appear as a result of work in this sector, researchers can continue to develop and improve existing solutions to the various problems to which they set out to achieve. Discussions such as Schnoebelen (Schnoebelen, 2017) and Leidner et. al (Leidner and Plachouras, 2017) already demonstrate the prioritisation of such considerations by making recommendations on best ethical practices to employ in future when solving various NLP problems. As technology advances, new problems within and without NLP will continue to appear. These problems will inevitably be accompanied by an increase in ethical considerations, moral, legal, or otherwise.

References

Alexander, L. and Moore, M. (2016). Deontological ethics. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition.

Beauchamp, T. L. and Childress, J. F. (2013). *Principles of Biomedical Ethics Seventh Edition*. Oxford University Press.

Berberich, N. and Diepold, K. (2018). The virtuous machine-old ethics for new technology? *arXiv preprint arXiv:1806.10322*.

Braunger, P., Hofmann, H., Werner, S., and Schmidt, M. (2016). A comparative analysis of crowdsourced natural language corpora for spoken dialog systems. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, pages 750–755.

Chowdhury, G. G. (2003). Natural language processing. *Annual review of information science and technology*, 37(1):51–89.

de Castilho, R. E., Dore, G., Margoni, T., Labropoulou, P., and Gurevych, I. (2018). A legal perspective on training models for natural language processing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, pages 1267–1274.

Driver, J. (2014). The history of utilitarianism. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2014 edition.

Escartín, C. P., Reijers, W., Lynn, T., Moorkens, J., Way, A., and Liu, C.-H. (2017). Ethical considerations in nlp shared tasks. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 66–73.

Fischer-Hübner, S., Hoofnagle, C. J., Krontiris, I., Rannenberg, K., Waidner, M., and Bowden, C. (2013). Online privacy–towards informational self-determination on the internet. *Digital Enlightenment Yearbook*.

Floridi, L. (1999). Information ethics: On the philosophical foundation of computer ethics. *Ethics and information technology*, 1(1):33–52.

Fokkens, A., Ruigrok, N., Beukeboom, C., Sarah, G., and Van Attveldt, W. (2018). Studying muslim stereotyping through microportrait extraction. In

Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018), pages 3734–3741.

Foot, P. (1985). Utilitarianism and the virtues. *Mind*, 94(374):196–209.

Harris, M. J. (2008). Consequentialism, deontologism, and the case of sheva ben bikhri. *The Torah u-Madda Journal*, 15:68–94.

Hellrich, J., Matthies, F., Faessler, E., and Hahn, U. (2015). Sharing models and tools for processing german clinical texts. In *MIE*, pages 734–738.

Hintze, M. and LaFever, G. (2017). Meeting upcoming gdpr requirements while maximizing the full value of data analytics.

Hursthouse, R. (1999). *On virtue ethics*. Oxford: Oxford University Press.

Johnson, D. G. (2004). Computer ethics. *The Blackwell guide to the philosophy of computing and information*, pages 65–75.

Koehn, D. (1995). A role for virtue ethics in the analysis of business practice. *Business Ethics Quarterly*, pages 533–539.

Koolen, C. and van Cranenburgh, A. (2017). These are not the stereotypes you are looking for: Bias and fairness in authorial gender attribution. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 12–22.

Leidner, J. L. and Plachouras, V. (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.

Lohr, C., Buechel, S., and Hahn, U. (2018). Sharing copies of synthetic clinical corpora without physical distribution—a case study to get around iprs and privacy constraints featuring the german jsyncc corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, pages 1259–1266.

Lyding, V., Nicolas, L., Bédi, B., and Fort, K. (2018). Introducing the european network for combining language learning and crowdsourcing techniques (enetcollect). *Future-proof CALL: language learning as exploration and encounters–short papers from EUROCALL 2018*, page 179.

Madnani, N., Loukina, A., von Davier, A., Burstein, J., and Cahill, A. (2017). Building better open-source tools to support fairness in automated scoring. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 41–52.

Manishina, E., Jabaian, B., Huet, S., and Lefèvre, F. (2016). Automatic corpus extension for data-driven natural language generation. In *10th International Conference on Language Resources and Evaluation (LREC)*, pages 3624–3631.

Melé, D. (2009). Integrating personalism into virtue-based business ethics: The personalist and the common good principles. *Journal of Business Ethics*, 88(1):227–244.

Mieskes, M. (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.

Mieskes, M. and Stiegelmayr, A. (2018). Preparing data from psychotherapy for natural language processing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, pages 2896–2903.

Molnár, G. and Szüts, Z. (2018). The role of chatbots in formal education. In *2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY)*, pages 000197–000202. IEEE.

Moor, J. H. (1985). What is computer ethics? *Metaphilosophy*, 16(4):266–275.

Plaisance, P. L. (2007). Transparency: An assessment of the kantian roots of a key element in media ethics practice. *Journal of Mass Media Ethics*, 22(2-3):187–207.

Rachels, J. (2012). Ethical egoism. *Ethical Theory: An Anthology*, 14:193.

Scheffler, S. (1988). *Consequentialism and its Critics*. Oxford University Press on Demand.

Schnoebelen, T. (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.

Sinnott-Armstrong, W. (2015). Consequentialism. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2015 edition.

Šuster, S., Tulkens, S., and Daelemans, W. (2017). A short review of ethical challenges in clinical natural language processing. *arXiv preprint arXiv:1703.10090*.

Suzumura, K. and Xu, Y. (2001). Characterizations of consequentialism and nonconsequentialism. *Journal of Economic Theory*, 101(2):423–436.

Tatman, R. (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.

Van Staveren, I. (2007). Beyond utilitarianism and deontology: Ethics in economics. *Review of Political Economy*, 19(1):21–35.

Voigt, R., Jurgens, D., Prabhakaran, V., Jurafsky, D., and Tsvetkov, Y. (2018). Rtgender: A corpus for studying differential responses to gender. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, pages 2814–2820.