Socio-economic landscape of digital transformation & public NLP systems: A critical review

Satyam Mohla† *Member, IEEE,* Anupam Guha

***Abstract*—The current wave of digital transformation has spurred digitisation reforms and has led to prodigious development of AI & NLP systems, with several of them entering the public domain. There is a perception that these systems have a non trivial impact on society but there is a dearth of literature in critical AI on what are the kinds of these systems and how do they operate. This paper constructs a broad taxonomy of NLP systems which impact or are impacted by the "public" and provides a concrete analyses via various instrumental and normative lenses on the socio-technical nature of these systems. This paper categorises thirty examples of these systems into seven families, namely; finance, customer service, policy making, education, healthcare, law, and security, based on their public use cases. It then critically analyses these applications, first the priors and assumptions they are based on, then their mechanisms, possible methods of data collection, the models and error functions used, etc. This paper further delves into exploring the socio-economic and political contexts in which these families of systems are generally used and their potential impact on the same, and the function creep of these systems. It provides commentary on the potential long-term downstream impact of these systems on communities which use them. Aside from providing a birds eye view of what exists our in depth analysis provides insights on what is lacking in the current discourse on NLP in particular and critical AI in general, proposes additions to the current framework of analysis, provides recommendations future research direction, and highlights the need to importance of exploring the social in this socio-technical system.**

***Index Terms*—Natural Language Processing, Digital Transformation, Artificial Intelligence, Socio-Political Impact, Business**

I. Introduction

In the current landscape of digital transformation, Natural Language Processing (NLP) has risen to one of the most important sub areas of AI/ML research and has a significant impact on not just the computer sciences, but linguistics, social sciences, and economics as well, the last due to its now influence on how labour is done. Communication via language is one of the most fundamental trait of human communities throughout human historical development [1] and how language is used has had always impacted socio-economic relationships in all human societies. Machines with the ability to use and generate human language (or "natural language" as it is called in the field) are changing and may further change multiple social and economic activities which were only performed by humans until recently. From translation of texts [2] to copy editing to interactive conversations, question answering [3], and eventually multi-modal fluent conversation [4] the current ability and future goals of NLP are having and will have a significant impact on how humans work, how they interact with machine intelligent agents and eventually how they interact with each other. This may also influence and alter extant power structures of human societies.

There is a growing realisation in the field of the ethical AI to decide what kind of NLP systems ought be researched and which should be completely avoided [5], [6]. There is also an increasing realisation that these decisions are not to be relegated to the dimension of only ethics [7] as there is a long tradition of scholarship which posits that social, economic, and political relations which decides the arc of scientific progress [8] are not captured by the ethical lens. It must be kept in mind that alterations to human labour, society, economics, and politics etc. are not deterministic changes [9] due to technological progress (in this case in NLP) as technological progress and framing of research problems does not happen in a vacuum, instead technological development including automation is a result of a series of political and economic choices by individuals, contending classes [10], and communities with differing interests. As such, this paper will attempt to not make an external trans-historical/moral critique of NLP but rather an "immanent critique" [11] which broadly reviews the world of NLP research and then examines its impact while proposing additions and alterations to the current framework of how AI in general and NLP in particular is evaluated.

The first contribution of this paper is a broad review into what are the kind of areas where SoTA NLP research has impacted the "real world" in the last five years. Seven such domains are examined and illustrated by 30 specific works. This is useful for anyone who wishes for a bird's eye view of research in applied NLP. The second contribution of this paper is to then analyse these seven interconnected families of research via instrumental, normative, and social lenses. The paper thus examines how is NLP broadly researched, designed, and deployed, and for whose benefits the field operates and also in a limited fashion examines on how that research gets impetus in the first place.The paper attempts to frame a future direction of research to measure long term impact of these NLP systems. We are aware that in doing so the paper itself alters the field of NLP ethics because it is making a decision on what kind of NLP systems deserve examination, commentary, and criticism, that is a weakness we acknowledge and hope our suggested future work complements.

† Corresponding Author {satyammohla@gmail.com}

Manuscript received Sept 02, 2022

Satyam Mohla is currently at Honda Innovation Lab, Tokyo and was affiliated with IIT Bombay, India & Koloro Labs. This work was supported in part by TechForSociety initiative at Koloro Labs & by UNESCO-LSUE Ethics in a Global Context Fellowship, University of Lucerne.

Anupam Guha is at IIT Bombay, India.

II. Related Works

While the field of AI ethics has matured over the last few years, NLP specifically has lacked as much as of a specific analysis, and yet there has been real progress in some areas. First, bias in both datasets and (generally deep learning) systems have been widely investigated and acknowledged [12] and multiple proposals like data statements and data sheets [13]–[15] have been proposed to correct for bias in NLP datasets. There has also been an acknowledgement on the community's part of dual-use [5] when OpenAI's GPT-2 was not initially released as a model over fears of abuse but finally was released after lack of evidence thereof and pressure from the research community. Thirdly, there has been attention drawn towards the real-world dangers of large scale use of language models, imputing meaning to what has been generated where none exists, and the connected financial and ecological impact of these systems [6]. There has been, in the area of computer vision, criticism of the pseudo-scientific nature of affective computing [16], both in terms of trying to measure psychological phenomenon from external physiognomic cues and also these arbitrary systems being used to police people [17] and perpetuate bias. We do a similar critique of affective computing in the context of NLP. Finally, new emerging works [18]–[20] examine the ethics of crowd-workers used in NLP research and make a case as to why "fair pay" only scratches the surface of what the community needs to consider while using these platforms to collect data. This last work opens the door to the least talked aspect of NLP research, the consideration of erased labour which creates the data NLP learns on [21]. These directions of analysis inform our work which is an attempt at gauging the landscape of what NLP systems exist in the "real world" and what are their broad impacts, social and economic. We hope our review work of a broad array of such systems contributes to the field of critical AI as well as informs NLP researchers.

III. Systems being reviewed

Our methodology consisted of two stages. First, to develop an analytical framework to identify avenues of socio-economic analysis missing from current discourse, a review was conducted. These avenues, namely instrumental analysis, normative analysis, function creep, political-economic impact & long term impact are discussed further in Section 4.

Second, a scoping review of documents, webpages and grey literature was conducted to capture ensuing changes across domains. We consider thirty such examples of NLP systems across seven domains. For each domains we will present these representative examples of work in the last five years which already have been deployed or have the potential to be deployed outside of the lab.

A. Financial decision making

The first family of NLP systems center around those applications aiming to expand the accessibility of finance by "judging" individuals. A set of problems in this family is concerned with the evaluation of creditworthiness, which complements and augments the traditional systems which use statistical methods on financial information with the use of machine learning methods. Another set of problems in this family is the use of NLP to identify who is likely to default on loans, essentially judging the trustworthiness of clients, and the third set of problems use similar NLP methods for various applications in the insurance industry, in the areas of marketing, underwriting, claims, reserving, and preserving. Illustrative examples in this family are:

1. SCOR SE reinsurance utilises NLP techniques for various purposes in this category [22]. They influence their insurance policy making and company strategy by using NLP to analyse commentary on social media websites like Twitter and Reddit. They are working on using NLP to extract information from medical reports to help underwriters focus on difficult cases. They are also working on developing NLP tools to analyse and classify claims as well as fraud detection. They propose to use these techniques to anticipate claim developments and expected costs.

2. Netzer et al [23], presents NLP techniques by which they claim that textual information from loan applicants can be used to predict their potential to default beyond the current financial and demographic information used to currently judge applicants. Their work further explores the potential traits of borrowers and claims to make a match between defaulting loan requests and the writing styles of extroverts and liars.

3. Crouspeyre et al [24], have advocated use of NLP techniques to augment existing creditworthiness measuring techniques based on financial information, for example FICO scores, by using machine learning approaches using non-financial information like phone log analysis and social media analysis. They claim NLP approaches will be less invasive than the former, lead to inclusion of more people in the financial system who do not have banking information history, and NLP techniques will be able to measure fraud via incoherency detection or a borrower's business knowledge etc.

B. Customer Service

The second family pertains to those NLP systems used by customer service industry to automate the customer service process and to analyse the feedback of customer service. These technologies include using sentiment analysis to evaluate how calls were handled by customer service representatives, or the use of NLP techniques to analyse large amounts of text from social media or message boards, or similar fora to analyse what is being said in public chatter, or the use of NLP systems to create bots and virtual agents to navigate complex customer service interactions etc. Illustrative examples in this category are:

4. Jia [25], sets out to do sentiment analysis on textual as well as acoustic data from a dataset of customer service calls, annotating the dataset for positive and negative sentiment. In this work they release a dataset of annotated service calls with an aim to fuse mutlimodal features like textual and acoustic data. They further attempt to do the annotation in a semi supervised manner, in order to get a lot of annotation done, and test multiple models to see which classifies the signals the best on their feature set. Using those features and models the authors claim a call can be given a numerical sentiment score.

5. Vermeer et al [26], attempts to discover the efficacy of various NLP techniques like sentiment analysis, dictionary-based methods, and supervised machine learning approaches, to categorise Electronic Word of Mouth or chatter about a company on social media, online forums, etc in order of how relevant they are for the company to respond to. They discover that content-based machine learning methods largely outperform sentiment analysis based methods to find what online chatter a company should focus on.

6. The Government Technology Agency of Singapore has been experimenting with virtual assistants (chatbots) [27] and according to their claims this has helped their citizens and businesses to significantly shorten wait times to queries, increased accessibility and in general improved user experience. Chatbot mediated public services are becoming increasingly sophisticated in Singapore as well as many other countries, and there have been attempts, for example DigiMo by Niculescu et al [28], to use sequence to sequence deep neural networks which learn on dialogue data focussed on a particular demographic (in this case Singaporean citizens), taking into account the emotional content in that data.

C. Policymaking and State

The third family of NLP systems being investigated are those being used in governance and to influence, inform, or build policy interventions by state in various areas. NLP systems are being used to evaluate large number of text corpora to rank innovation mechanisms, to investigate media texts to isolate incidents across the globe to detect triggers of emergencies, evaluate public sentiment via text and so influence policy in so called “smart cities”, for selecting relevant texts and organising knowledge bases from a large amount of scientific and technical documents so that policymakers can easily identify which ones to study. Illustrative examples in this category are:

7. Lin et al [29] working for the US Department of Energy’s Oak Ridge national laboratory have developed a methodology using NLP that transforms text and numeric data into a geographic mapping and detection of where clean energy innovation is happening. Using this tool which detects, measures, and characterises clean energy ecosystem innovations, one can help reach policy decisions knowing what kind of work exists regionally etc.

8. Xiong et al [30], build a case study for using social media posts to mine public opinion for informing crisis management policy during emerging environmental threats. They use LDA to identify topics discussed on Twitter about the 2019 Chennai water crisis in India. They classify tweets on this topic and examine the relationship between the chatter on the social media platform and rainfall precipitation levels.

9. Alam et al [31], built a novel method for social media sentiment analysis to use them in smart cities application, they interpret multiple hyper parameter combinations for neural networks to find which work best for datasets obtained from Twitter.

10. Pérez-Fernández et al [32], built CorpusViewer which helps policymakers analyse various documents before they are studied like patents, scientific publications and public aids for gathering evidence for policy implementation. It automatically classifies documents into a taxonomy, does basic topic modelling on them, provides document similarity and plagiarism detection, semantic area detection, temporal analysis and other analysis tools.

D. Education

Connected to the previous group of NLP systems, the fourth family of NLP systems the paper covers are being used in the area of education. There are a lot of use cases in education with NLP systems being used in pedagogy on language learning apps, they are used in MOOCs, bots being developed to automate parts of the teaching process, NLP systems are being used to track the evolution of language competence, they are also being developed to automatically score essays, and they are being used to critically evaluate essays as well. NLP systems are also being used to accelerate education is specialised areas. Illustrative examples in this category are:

11. Automatic essay evaluation is a large area of interest in this family. Generally, AES tasks are of three types, a regression task to predict the score of an essay given some metric, a ranking task to rank a series of essays according to quality, and a classification task to place essays into some quality categories. Bhatt et al [33], improve on previous methods to automatically evaluate essays by NLP methods. Their method primarily uses semantic similarity of sentences in addition to rule based grammar and consistency tests and attempts to replicate features which are considered by human graders.

12. Miaschi et al [34], develop a NLP method which uses computational stylometry to track the evolution of written language competence of Italian L1 learners. Their work tries to predict the chronological order of two essays written by the learner in different times, and they investigate which language phenomena aid in this prediction task and how that changes over time. They are focused on features modelling the form and not the content of the text.

13. Duolingo has done a significant amount of research work in NLP methods. Duolingo’s Second Language Acquisition Modeling (SLAM) task [35] involves a large dataset of beginner level student data of three different exercises to model how a student acquires and learns a new language. Duolingo’s Shared Task on Simultaneous Translation and Paraphrase for Language Education (STAPLE) [36] involves getting a set of translations for a given sentence into five sentences and unlike conventional machine translation tasks instead of comparing the result to one reference the set of proposed results is compared to a weighted set of references and a weighted F1 result is computed.

14. MOOCs have become popular which has led to a demand for automating parts of the pedagogic process as a lot of students are involved in these. One particular area is predicting from student forum posts where an instructor intervention is urgently needed. Alrajhi et al [37], combine NLP methods with a deep learning model to classify such form posts.

15. Poce et al [38], have worked on discovering which NLP features extracted from Italian text reveal the most about

critical thinking of the essay writers, which is measured by using six critical thinking sub dimensions assessed by humans. This and related works are used for automatic evaluation of critical thinking in essays written by students.

E. Healthcare

The fifth area of NLP systems we investigate are those being used for medical applications and healthcare. These usages cover a lot of different areas with the use of NLP systems by epidemiologists to flag spread of infectious diseases by going through large numbers of worldwide media reports, using NLP on flight data to figure disease flashing points, using NLP systems to generate medical reports from patient doctor conversations, using NLP systems to analyse texts to detect diseases like early stage Alzheimer, using NLP systems to review scientific literature to accelerate drug discovery processes, and using NLP systems in telemedicine, i.e. remote interactions of doctors and patients mediated via chatbots to automate large parts of the interaction. NLP systems are now also being used to provide advice in mental health interviews of patients with chatbots. Illustrative examples in this category are:

16. Canadian firm BlueDot [39] gathers text data from about a hundred thousand articles in more than 65 languages everyday in addition to airline ticket data globally, as well as censuses, climate data, publicly available statistics reports, global infectious diseases alerts, etc to find disease hot spots, and can send alerts to its clients. BlueDot's automated infectious disease surveillance proved effective during the 2009 H1N1 influenza outbreak, the 2014 ebola outbreak, and was able to predict the outbreak and the high risk cities for covid-19 in 2020 nine days before the USCDC and WHO sent out notices.

17. Enarvi et al [40], explore NLP methods to automatically generate medical reports from transcripts of patient doctor conversation. They compare two methods, a hierarchical RNN with a pointer generator network and a transformer based sequence to sequence architecture.

18. Alzheimer's disease accounts for 60% of dementia cases and early diagnosis is extremely useful for treatment and management of the disease. NLP techniques have proven useful in early diagnosis of this disease, by detecting symptoms such as Mild Cognitive Impairment (MCI). Li et al [41], have created a dataset called B-SHARP of speech transcripts on various topics which can be used to detect MCI and have explored methods for the same using transformer encoders.

19. Drug discovery is a process which involves discovering protein targets using the principles of how certain compounds interact with protein. This process can be automated with NLP methods as text based representation of biochemical entities are relatively easily available. Ozturk et al [42], explore the use of NLP methods to analyse the text based representations of chemical compounds to accelerate drug discovery.

20. The covid-19 pandemic has seen a sharp rise in usage of chatbots in various healthcare usecases and a particular use is that of using it in digital medical health. Wysa [43], a mental health chatbot has more than a million and a half users and uses Cognitive Behavioural Therapy (CBT). Another such chatbot app is Woebot [44] which aside from using CBT methods also uses a survey to tailor itself for the user. Meadows et al [45], analyses and compare these two systems and find that while Wysa needs more information and thus asks more open ended questions, Woebot is a bit more scripted. Neither system can replace actual therapists.

F. Law

The sixth family of NLP systems the paper covers are those used in law. These NLP systems are used in providing legal advice via chatbots, discovering information relevant to a case to make better decisions, going over a contract to make sure it is complete, eDiscovery that is finding how relevant is a document to case, and similarly generating documents for a legal case. Illustrative examples in this category are:

21. In their work on legal judgement prediction, Yang et al [46] construct a dataset of the same kind of charge with trial information as well as information on the attitude of the suspect and create a model to discover the relationship between the suspects attitude and the penalty of the case demonstrating that there is a relationship between the two.

22. Ruggeri et al [47] use Memory-Augmented Neural Networks (MANNs) to expose unfairness in legal contracts by generate useful explanations, this is done by training a MANN on a corpus of online Terms of Service which can then detect unfair legal clauses as well as provide possible rationales behind those. The authors have evaluated multiple MANN configurations and improved classification and explainability from previous such works, aside from providing this dataset.

23. Queudot et al [48], create two function specific legal chatbots, the first to provide information on immigration issues using data from the Government of Canada, and the second to provide information on job related legal issues to bank employees. Both these chatbots are trained on FAQ data, the former on online data obtained from the Government of Canada's Immigration and Citizenship Help Desk and the latter on a FAQ within the bank. The chatbots use the standard RASA model.

24. Sugathadasa et al [49], create a mechanism for legal document retrieval by combining two methods for document vector representation. Their work demonstrates the utility of incorporating semantic similarity measures into such IR tasks for domain specific documents.

25. Summarising legal documents automatically has been an area of research. Jain et al [50], use a Bayesian optimisation approach to tune the hyperparameters of a classical text summarisation algorithm called Textrank by optimising an objective function based on the ROUGE score. This lets the relatively simple algorithm like Textrank perform well in summarising legal documents.

G. Security

Connected to law, the final seventh family of NLP system this paper covers are those being used in policing, surveillance, defense, and national security. NLP systems are being used to detect hate speech, investigate radicalisation of individuals and communities, and are being used to supplement other AI mechanisms for the use of predictive policing. NLP systems are being used in automating surveillance of online communities as well as individuals, and are being also used to refine metadata. Illustrative examples in this category are:

26. Alshehri et al [51], create a model to detect dangerous speech in Arabic Twitter, namely intentional threats. They focus on physical harm threats. They create a dictionary of multi dialectical physical harm threats in Arabic and collect a large dataset of threat data. They manually annotate a section of their dataset and analyse what kinds of threats are usually made. They train BERT variants with Arabic emotion data to detect these texts.

27. Araque et al [52], use insights from affective computation (that is computational methods to detect and process human emotions) and a resource called the SenticNet which is a knowledge base of concept level sentiment analysis to extract two feature extraction methods which they use to improve their classification performance on hate speech detection tasks. In a previous work on classifying radical text against neutral text or anti-radical text, Araque et al [52] use an emotion lexicon which has words annotated with emotions and a radical lexicon which uses word embeddings to measure the semantic similarity of a text with it. They find emotion features help the classification.

28. Percy et [53], analyse crime records of different regions and different time periods as text documents with NLP tools to predict crime patterns. They attempt to identify regions with similar patterns of crime and try to cluster regions with similar crime patterns over time.

29. Sun et al [54], use NLP methods to build a digital forensic investigation platform which can investigate criminality in online communities. The method needs a corpus of communications where ideally both the senders and receivers are known, or at least the sender in case of social media, it then uses topic modelling to discover what is being talked about, refines the topics, and the uses a set of classifiers using the topics as features to detect criminality of the participant.

30. Ziems et al [55], work on cyberbullying detection creates a new annotation framework and provide a new dataset based from Twitter data to detect cyberbullying and distinguish it from other kinds of online aggressive language which are detected by existing classifiers which can't distinguish cyberbullying specifically.

IV. Analysis

In this section we will cluster aspects of these families, instrumental and otherwise and look at them critically through a variety of lenses. While we discuss the potential weaknesses and unintended real life impacts which may be harmful of these systems or their variants, we are not discounting their potential benefits which have been motivated by their creators. Indeed, within the very tight constraints these systems ought to be used in the benefits usually outweigh the harm, and most of the harm is in using these systems blindly and outside the very specific contexts they are designed for, something we discuss in the sub-section on function creep.

A. *Instrumental analysis*

The first and the most direct way of looking at an NLP system is to evaluate the mechanism of a particular algorithm in a particular usage. In general, the question of how datasets are collected, what their priors and assumptions are is essential to evaluate the strengths and weaknesses of these NLP systems, in which contexts do they work and where they prove to be brittle. One can look at how much fairness, accountability, transparency (FAT) is possible in these algorithms and if their data or algorithms can be potentially biased. Then one may consider the strengths and weaknesses of the specific models used in the above use cases. With NLP systems the more well defined the task is the easier it is to predict how it will fail.

Let us take examples 2 and 3. These require datasets where there is a connect between the "nature" of humans and how they use language. One weakness of dataset collection arises here in semantic heavy NLP tasks is the unsettled nature of annotation agreement. While a set of annotators which fluency in a language will always accurately judge which parts of a text to annotate for a coreference resolution task or a question answering task, it would be harder for them to come to an agreement for annotating a text for an irony/sarcasm detection task, and even harder still to come to an agreement annotating something like aggression, or the artistic merit of poetry, or the humour content in a joke embedded in a piece of text. This is because while for some artefacts of language there is a one to one well defined correlation between mental "internal states" or semantic content and the way it is written in natural language, with others these mappings become both vague and distributed over a piece of text and when we approach affective computing related concepts like emotions and humour, there is disagreements among psychologists whether things like universally articulated emotions even exist. On the far end of the spectrum some of the claims to map internal emotions etc with any degree of universal certainty with patterns of text is unfalsifiable and pseudoscientific.

For example system 4 claims to have a strong understanding of sentiment, while system 11 attempts to find the "quality" of a written essay. Such use cases are prone to arbitrariness even among humans. To compound this arbitrariness of datasets, a trivial but oft forgotten instrumental aspect of NLP when used in off shelf applications is that no matter how good the model, it is probabilistic in nature, and with many models it is also not easy to determine why an error happens when it does, thus when models fail, they fail gracelessly, that is their incorrect results are also arbitrary. While there is an ever-increasing area of research to bake in explainability in NLP systems this is still nowhere near the standards human policymaking needs. This has real world consequences on use cases which are implicated in policy, governance, justice, financial decisions, and security mechanisms which by definition must not be arbitrary. Not that NLP techniques cannot be used at all in these areas, but their uses should be limited to provide insights with strong checks and balances with human oversight, they must not be used to provide decisions themselves. In these use cases there is an ongoing debate in literature on what are the hard limits of possible accountability and transparency.

As we see in the systems 2 and 3, there are deeply invasive decisions being made on customers to predict creditworthiness and future bad behaviour, and these decisions can be both wrong and arbitrary. The fact that an algorithm made this decision also lends a false sense of credibility to them, and as has been seen in studies like cite study employees tend to pass responsibilities to opaque AI systems when there is no way to tell why a customer has been rejected. Examples like system 26 and system 27 if used in real world use cases can have multiple levels where they can go wrong (the data itself, the annotations not accounting for all kinds of contexts, the model being overfit on the data, and an interpretation of the mode's results) and very easily break down in linguistics contexts they are not designed and tested in. Decision making using NLP tools in domain involving social and economic rights of people must be avoided, as the decisions may be incorrect, arbitrary, and non-transparent. Moreover, while a human agent in a similar application can also be wrong or biased, one can fix moral and legal responsibility on human agents using laws and regulations, which do not apply for algorithms.

Another application is a family of algorithms which decide what is useful out of a collection, thus using NLP techniques of recommending instances out of a data collection. Recommendation algorithms can have various benign uses in media websites, e commerce, and policymaking (see system 10) but some variants are now being pushed in law and judiciary use cases like system 24. The strength of these techniques is to massively accelerate document discovery process which being a lower end task full of manual drudgery is attractive to legal practitioners to decrease the delay of legal work and thus increase accessibility of legal resources which are scarce for the general citizenry of many countries. A variant of this use case is to use NLP document retrieval to fill templates and provide automated legal advice, similar to system 6. But there is a potential area of concern in this, as recommendation systems imply that whatever document is deemed useless will never reach the human observer for consideration. This is not usually a problem in media or e commerce use cases because the stakes of not finding the "correct product" are not close to what can happen in a legal environment. NLP for document discovery while useful to accelerate the legal process must not be used in cases where missing documents might irrevocably harm the life, rights, and liberty of the legal participant, in any usage of such a service must carry the alternative of human led discovery. In applications where there is dearth of humans to discover documents and yet no peril of a missing discovery causing harm, like scientific research, art, media, and scholarship, these methods are extremely useful.

When NLP tools are used to provide insights rather than decisions, often unsupervised learning tools like clustering and topic modelling is used to group documents, discover structures and hierarchies to automatically organise them. Examples include systems 7, 8, 9, 10, which use variants of topic modelling and as long it is understood that such unsupervised techniques only display statistical aggregates and not specific answers, they can be safely used to get insights from.

B. Normative analysis

The overarching weakness of instrumental analysis (and thus the FAT framework) is that it is agnostic of ethical and human rights implications of these systems and ignorant of what values socially and politically these systems might be promoting or preventing. For example, while system 11 might have technical flaws in how representative its dataset is etc the prime concern with it is that even with an excellent dataset and tested algorithm this is stochastic, and one does not want arbitrary exam results without human accountability in case the exam has a significant impact on the student's career. This risk is not completely balanced by any benefit of speed or scale and thus any use of this system deserves scrutiny. Being used in the real world none of these systems are socially and politically neutral, in fact several of them make assumptions about what is desirable in the world without explicitly stating that in so many words and thus run the danger of replicating the past, at times a past which should not be replicated. We are however cautious of overstating the case for a normative analysis of these systems as we realise that not only is there a lack of a shared understanding of ethics between the plethora of AI ethics councils, there is the further problem of a lack of accountability and redressal mechanisms for violation of ethical concerns. Moreover, due to the above the discourse on ethics devolves into what some observers call ethics washing, that is a cynical use of the language of ethics by companies and states while making no concrete commitments. To make our normative analysis of these systems more concrete we would augment the ethical lens by grounding it in human rights as suggested by Marda [56]. Human rights have a more coherent enforcement mechanism and internationally recognised legacy. The guiding principles of business and human rights for example instruct states to protect human rights in their territories even if the violations are done by non-state entities like businesses and that covers developers and customers of AI systems.

Among the systems listed, system 28 is one which predicts preponderance of crime in a region. As has been observed in the past, predictive policing has a habit of "predicting crime" in overpoliced regions which are inhabited by the poor and racial minorities by replicating patterns of police behaviour. Any attempt to find patterns of crime merely by past patterns runs the risk of overdetermining what may happen in the future, and does influence policy of state repression. In a scenario where these tools often are used without realising their inherent limitations, this use should be argued as potentially unethical as it attaches the label of criminality to the residences of an area which is collective guilt, an unethical authoritarian policy prior. System 20 uses chatbots to provide a semblance of mental health care. While just chatbots do give disclaimers that they are not to be used as actual practitioners and are not replacement for therapist, what is missing is the recognition of the phenomenon that people do assign moral agency to such chatbot systems (ever since the infamous ELIZA experiments) [57], [58] and it is unlikely they will not take seriously a mechanism which at the end of the day is a stochastic parrot. To provide medical care from a statistical care is unethical as it violates the basic requirement of moral agency which a medical practitioner needs to have. Systems 1, 2, and 3 extend the logic of current credit scoring systems and use NLP to make strong inferences about the character of individuals. These systems aside from this being technically weak due to reasons of arbitrariness and impossibility of deterministic outcomes, where they differ from say algorithms which just use financial data, they are also ethically suspect because of the potential of economic harm on people for no action but a presumption of intent.

The potential for unethical use of systems is the most in the last category, namely security, because the carceral and security apparatus of various countries get exceptions in both data security and human rights mechanisms and also have a history of solutionism and bad usage of digital systems. For example, systems 26 and 27 which must not be used without human oversight can be used by law enforcement agencies and there is a historical precedence of usage of ML systems, especially computer vision systems like Facial Recognition Technology (FRT) being used injudiciously. Human oversight is also not a guarantee of weeding out false positives as we see in attempts by various platforms to introduce human moderators to verify contents ML algorithms mark as hate speech etc. What is often seen with these companies that these moderators are not hired as much as needed due to confidence in the algorithms, thus often overworked, and being continuously exposed to violent content is deleterious for their mental health. As mentioned before, the presence of a machine giving an insight is often given more gravity than a human giving it in security settings and thus can give legitimacy to otherwise clear violations of human rights and natural justice. The potential harm of these systems is amplified by the gap of technical literacy between the researcher and the policymaker leading to these systems being used in a way they were not meant to which leads to function creep.

C. Function creep

Function creep refers to the steady expansion of use cases for a given entity beyond its initial planned use. Originally this term was used for how data and datasets collected for ostensibly harmless and beneficial purposes end up being used in many more areas without the consent of the people whose data was collected for other stated reasons. In the lack of data protection legislation in most countries this phenomenon is common. Here, we speculate about the potential function creep for these NLP methods given past instances of how this happens. One famous example of function creep is that of the US algorithmic tool used to predict criminal recidivism called COMPAS an acronym for Correctional Offender Management Profiling for Alternative Sanctions [59]. COMPAS was originally designed not for sentencing but to predict what kind of help a convicted inmate might need, for example mental health treatment, to make their reform easier. Then its use was extended into making a decision on what should be the condition for releasing convicts and under trials, for example whether to grant them bail or not. This jump was done without recognising that no algorithm should be given that level of control over a human's liberty. And currently its use has expanded in multiple US jurisdictions for criminal sentencing. This is also despite the developers of the COMPAS system not initially designing it to be used for sentencing or being confident of its use in sentencing. COMPAS has come under criticism on accounts of being biased against racial minorities however there is the question of whether such systems be used at all, bias or no, given that human rights scholars are in universal agreement that decisions on people's right to liberty cannot be based on assumptions of what they might do.

Among the representative systems digital forensics datasets like those used in systems 29 and 30 are very useful resources to do research on how to regulate platforms, and discover patterns of hate speech and threats they amplify, but can very easily turn from being tools of research to tolls directly used by law enforcement who may not grasp the nuance that these systems need human verification regardless of their error margins. Chatbot mediated public services like system 6 are at the surface a benign and useful application which improves accessibility and speed, but they should not be used in places where the lack of a human official could cause harm to a petitioner who might need someone to interact and discuss a complex problem with. Similarly, the use of chatbots in telemedicine could give to spurious use cases like system 20, or simply divorce the aspect of responsibility of a healthcare provider from healthcare advice while maintaining a legal immunity. In such applications at the minimum there should be the option to interact with a human. Topic models like system 8, 9, and 10 are excellent tools to get high level insights for policymaking as long as it is understood said insights are quite open to interpretation and their use should not creep towards actual framing of decisions without sufficient due diligence of collecting pertinent facts on the ground. System 17 is an automated summarisation mechanism for medical interviews between patients and doctors, and this has the potential to be given an expanded role in places where there is a dearth of medical practitioners, and thus must never be used in applications in which missing out on some pertinent fact in a summary could result in actual harm to the patient. System 11 which is developed for automatic essay evaluation has the potential to be used not just in controlled pedagogical spaces with human oversight but also creep into self-teaching via apps where it is easy to confuse speed and convenient for actual teaching potential. This kind of harm is hard to detect because of trust people have in technology and its perceived sterility. Also, a lot of function creep is pushed due to both the existence of the profit motive on the part of AI vendors pushing them towards not stating the many weaknesses of contemporary NLP systems, a desire of businesses for a lack of regulations on technology, as well as solutionism on the parts of state officials and policymakers who are bent on retreating from the policy space. This leads us to a political economic analysis for these systems.

D. Political economic analysis

Political economy refers to charting the relationship between what is being studied, in this case these categories of

NLP systems, and the inter relations of society and state. We are primarily interested in how these NLP systems impact and are impacted by socio-economic relationships, governments, and public policy. Let us first look at how they effect the mode of production, which means how overall a society produces goods and services. When analysing these NLP systems from this particular lens we first have to look at how labour is done to create these systems and how they impact labour. A significant part of the labour for NLP systems of all of these categories is the production and/or collection of data. In some of these categories, like finance and customer service, and data production which involves, collection, cleaning, structuring, and annotation is in-house, done by private companies or government organisations. In some examples however, like systems 8, 9, and 10 the data is obtained from the public domain, by either scraping social media or other repositories of public data. While regulations on use of public data for commercial and policy making differs from country to country (in Europe for example GDPR prohibits certain collections and use of data), there is a larger debate on “community data”, i.e. data obtained from communities of people to be used for the benefit of those communities. Some data, though publicly obtained, are not related to or produced by people, like weather data, and such impersonal data can be freely used. Another aspect of data related labour in these systems is when the data is collected via crowdwork platforms like Amazon Turk or crowdflower. The proliferation of these projects promotes platform work. There are ongoing debates on the ethics of using workers using these systems as work on them are generally not regulated by labour laws, and often these platforms insist that the workers they employ are not workers but contractual associates. Also, this implies that jurisdictions which have robust labour laws but no regulations on crowdwork platforms can inadvertently incentivise shift to some NLP use cases (like reducing medical practitioners and instead using 17 or 18, or 19 instead of manual drug discovery, promoting legal AI systems like example 23 instead of hiring more employees) because it is possible to legally underpay crowdworkers who can be used to create the datasets running the NLP systems designed to replace certain roles. The argument here is that these systems are not necessarily doing work faster or better, but because their data collection is “hidden” and it is easier to exploit crowdworkers, they yield more profit and promote further platformisation.

The use of such NLP systems results in a general precarity and wage depression of workers in that field, and also this precarity affects the marginalised communities more. Specifically, women who are twice exploited in the domestic sphere and at the workplace are impacted more. Additionally, the use of such systems in sensitive domains like healthcare and judiciary while attractive in a vulgar short term economist sense, it is actually harmful as all these systems are stochastic and will have errors, and will also erase official accountability as discussed before, and so the costs are externalised to the vulnerable in society.

As for their impact on labour, some of these systems like those covered in education and healthcare, alter the way work is being done in these fields. Automatic essay evaluation, NLP applications in language learning platforms, and NLP applications in MOOCs have the potential to alter the classroom and it is again likely to be pushed through if they reduce, or are perceived to reduce, the costs of hiring and training teachers. Again, this is not hypothetical but already happening in many countries, especially the global south. In a global economic climate of privatisation when states retreat from their welfare duties, these tasks like education, healthcare, and even aspects of judiciary and policy making are taken over by private companies which are have very different priors than democratic governments, and will find these NLP techniques labour saving and profitable. It is of concern that said alterations in how work is done in education is happening before educationists and pedagogy experts can figure out whether these could have deleterious effects on the students. In fact often the proliferation of these systems present a fait accompli to policy makers before impacts are analysed. In healthcare, example 17 and 20 are use cases of what could be pushed as a bandaid measure by states and companies if they are perceived to be adequately robust and the training and hiring medical workers is considered expensive. Again, the actual impact of stochastic (and therefore unintelligent) systems like these are not studied by the time the replacement of public policy by NLP infrastructure which is developed and maintained by private companies.

Thus we observe that there is a significant influence some of these systems have indirectly on not just governance and policy making, but their use by private companies potentially alters the framework of what is governance and policy making, steadily eroding the policy space and replaces it by a mechanism which by definition is stochastic and non accountable as all machine learning is stochastic and only humans have moral agency. More directly, as the sections on on policy making and legal tools, systems 7, 8, and 9, 22, 24, 25 demonstrate there are off shelf systems which are altering the way these areas work. Specifically, tools like system 24 and 25 can harmfully alter how burden of proof works diminishing accountability (by erasing which document and what text in a document is relevant), while systems like 8 and 9 which use extremely stochastic techniques to inform decisions introduce more arbitrariness in policy making and governance.

E. Long Term Impacts

As the above lenses indicate, there is a potential long term impact on social relationships by the use of these various kinds of NLP tech in the public domain. These impact are not just on economics at a surface level but also a larger impact on how knowledge and culture is formed and shared, how inequity and hierarchy reify, etc. The section on instrumental analysis indicates that popular use cases exist where the decisions these systems come to cannot not be arbitrary in some sense, for example systems 2, 3, 26, 27, but what it misses out on is the unearned trust these often incorrect systems have among their human users who will have the bias to conflate machine results for rigor and also the apathy not to question it. Such behaviour has been observed in organisations where officials use machine learning to come to decisions neither have the incentive to question the decisions nor the knowledge that the decisions can be as wrong as human ones, and unlike human decisions are not subject to accountability, petitions, or negotiation. There is an economic incentive for bad actors for promoting this as a culture in the organisations they control, and hence just saying that these systems are imperfect will not have the desired effect without regulation. Similarly, in the normative analysis while the ethical pitfalls of using systems with no agency to make ethical decisions (most medical, legal, and policy decisions have an ethical component) has been pointed out, what it misses is the ability of these systems to erase that ethical question in the first place when presenting the problem in from of the user. For example, while a lawyer may be alarmed by the idea of NLP systems setting the charges on a crime report, they might not be if all the NLP system is doing is recommending which documents to read for a case. But even in a latter benign use case the NLP system has already decided for the human what not to read and what fact is irrelevant, which is a moral decision. Such an erasure is in many senses convenient in say an already overburdened legal systems incentivising quick "solutions" to clear its backlogs. Thus, while one may find immediate problems individually with these systems what gets ignored in research is their ability to make "convenient" complex problems of the public sphere and removing those problems from public discourse, oversight, and challenge and thus making the staus quo comfortable. This leads to a research direction where the goal is to "solve" things like bias and transparency in the sense of solving an optimisation problem, or create "better datasets" in order to engineer away the issues with bad or unsound knowledge which is the base of some of those use cases, and keeps researches distracted from the conclusion that for some of these systems, their public use must be thoroughly regulated if not outright prohibited, that one cannot use fundamentally stochastic systems on problems which should not be solved in a stochastic manner. What is a fundamentally policy or political problem cannot purely have a technical solution. Also, it ignores the very real problem that regardless of the pointing out everything from the bad datasets, flawed algorithms, potential for function creep, or a direct potential harm to rights of people, why NLP systems which should not exist in the public space do exist. They do because of the seeming convenience and hence concrete profit they can generate in mystifying social problems. Our work advocates future research in trying to understand and thus challenge this popularity.

V. Limitations and Suggested Research Direction

The previous analyses point to the pattern of NLP systems being used in the public sphere having serious flaws which exist across their categories. There is a tendency for data being based on shoddy priors, having incomplete and biased datasets, systems being by design non transparent, encouraging a lack of agency of their users, thoughtless repetition of past behaviours via machine learning, and substituting stochastic behaviour for intelligence. But it also points out the reasons these flaws exist are not just ethical gaps or flaws of design as much as they are products of genuine limits of what extant NLP is capable of outside of specific niches and data sources they are designed and developed in and their incorrect and overuse for social and economic reasons. Substituting human accountability, something needed in most public infrastructure, with seemingly intelligent language and a certain degree of shallow coherence will inevitably cause these problems and the only way to control that is rigorously limit and regulate the use of NLP systems in the public domain. While there are genuinely bad priors, like attempts to ascribe emotions to patterns of texts, a lot of these flaws are not purely technical flaws. Also, the social gaps these systems are made to ”solve” are genuine and until those gaps are addressed by correct policy making it will be incorrect to focus just on correcting NLP research. The limitations of our paper is that it looks at the various categories of NLP systems but not so much towards the people and organisations which use them for their benefits and despite their flaws. The ambition of the paper is purely descriptive, only to present a taxonomy/review of NLP systems used in the public domain and provide an in depth analysis of them, but future work is needed to do concrete policy recommendations for various jurisdictions, political economic realities, and policy regimes on how these systems ought to be used or not used judiciously. Additionally, as NLP is a family of socio-technical systems a sharper focus is needed on the communities, cultures, and especially capital which goes into the conception and production of these systems. Whether it be academia or the industry, flawed or brittle NLP systems are not designed or developed just because of a lack of ethics training of researchers, such a reading would be quite vulgar. Rather we posit that NLP research is both a product of and is implicated in patters of capital which incentivise an ersatz automation of language regardless of its flaws and we would like to investigate this in our future work. To NLP researchers and developers we urge that they should have a consideration for the impact of their research beyond the level of individual ethics and transparency, but also towards how their work might alter the nature of social and economic relationships. The impact of NLP research is beyond shaping the direction the field itself takes, outside the laboratory it has the potential to influence how value is produced and accrued in society, how knowledge and culture evolves, and how work and wage operate.

VI. Conclusion

Among the various fields of research in artificial intelligence, thanks to rapid developments in machine learning research as well as a jump in hardware capability, NLP has gained a lot of popularity in the last few years and have come into its own in the public domain. From private usage to market products to state policy making, we are observing a rich variety of NLP use cases. However, this development has quickly outpaced and left far behind scholars of society and policy in how these systems impact what is "public". There has been no work yet to provide a taxonomy of NLP systems in the public domain with an analysis of their social and economic impact. In this paper we give a broad and high level review of extant NLP systems in the public domain which have had tangible impact on the "real world" in the last five years, and we illustrate our categories with thirty example systems. We then give an analysis using various lenses of the features and flaws of these systems, their uses and potential abuses, and how they have a larger impact on society. We ask our readers to consider whether or not the research arcs of these systems are inevitable or are they products of certain incentives these systems or the political and economic milieus they operate in generate. If the status quo is not ideal, what should be the future arc of NLP research where it impacts the "public"?

We hope our work is useful to a wide array of scholars, from someone wishing to know a birds eye view of what NLP research has current public uses, to researchers and policy makers investigating the potential risks or unintended consequences of these systems. We also hope for a broader collaboration between NLP, policy, economics, and legal scholars to better explore the social side of these socio-technical systems.

Acknowledgements

The authors would like to thank the anonymous referees for their valuable comments and helpful suggestions.

References

[1] J. Hurford, "The evolution of language and languages," *The evolution of culture*, pp. 173–93, 1999.

[2] P. Koehn, *Statistical machine translation*. Cambridge University Press, 2009.

[3] B. McCann, N. S. Keskar, C. Xiong, and R. Socher, "The natural language decathlon: Multitask learning as question answering," *arXiv preprint arXiv:1806.08730*, 2018.

[4] G. Daniel, J. Cabot, L. Deruelle, and M. Derras, "Xatkit: a multimodal low-code chatbot development framework," *IEEE Access*, vol. 8, pp. 15 332–15 346, 2020.

[5] K. Leins, J. H. Lau, and T. Baldwin, "Give me convenience and give her death: Who should decide what uses of nlp are appropriate, and on what basis?" *arXiv preprint arXiv:2005.13213*, 2020.

[6] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 610–623.

[7] E. Bietti, "From ethics washing to ethics bashing: A view on tech ethics from within moral philosophy," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 210–219.

[8] S. Borrás and J. Edler, "The roles of the state in the governance of socio-technical systems' transformation," *Research Policy*, vol. 49, no. 5, p. 103971, 2020.

[9] J. Durham Peters, "You mean my whole fallacy is wrong: On technological determinism," *You Mean My Whole Fallacy is Wrong: On Technological Determinism*, pp. 26–34, 2019.

[10] F. N. David, *Forces of production: A social history of industrial automation*. Routledge, 2017.

[11] R. J. Antonio, "Immanent critique as the core of critical theory: Its origins and developments in hegel, marx and contemporary thought," *British journal of sociology*, pp. 330–345, 1981.

[12] I. Garrido-Muñoz, A. Montejo-Ráez, F. Martínez-Santiago, and L. A. Ureña-López, "A survey on bias in deep nlp," *Applied Sciences*, vol. 11, no. 7, p. 3184, 2021.

[13] E. M. Bender and B. Friedman, "Data statements for natural language processing: Toward mitigating system bias and enabling better science," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 587–604, 2018.

[14] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford, "Datasheets for datasets," *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021.

[15] T. Schick, S. Udupa, and H. Schütze, "Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1408–1424, 2021.

[16] L. Stark and J. Hutson, "Physiognomic artificial intelligence," *Available at SSRN 3927300*, 2021.

[17] V. Marda and S. Ahmed, "Emotional entanglement: China's emotion recognition market and its implications for human rights," 2021.

[18] B. Shmueli, J. Fell, S. Ray, and L.-W. Ku, "Beyond fair pay: Ethical implications of nlp crowdsourcing," *arXiv preprint arXiv:2104.10097*, 2021.

[19] S. Mohla, B. Bagh, and A. Guha, "A material lens to investigate the gendered impact of the ai industry," in *IJCAI 2021 Workshop on AI for Social Good*, 2021.

[20] J. K. Kummerfeld, "Quantifying and avoiding unfair qualification labour in crowdsourcing," *arXiv preprint arXiv:2105.12762*, 2021.

[21] M. L. Gray and S. Suri, *Ghost work: How to stop Silicon Valley from building a new global underclass*. Eamon Dolan Books, 2019.

[22] S. SCOR, "Technology and the future of reinsurance," 2020. [Online]. Available: https://www.scor.com/en/expert-views/technology-and-future-reinsurance

[23] O. Netzer, A. Lemaire, and M. Herzenstein, "When words sweat: Identifying signals for loan default in the text of loan applications," *Journal of Marketing Research*, vol. 56, no. 6, pp. 960–980, 2019.

[24] C. Crouspeyre, E. Alesi, and K. Lespinasse, "From creditworthiness to trustworthiness with alternative nlp/nlu approaches," in *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*, 2019, pp. 96–98.

[25] Y. Jia and S. SungChu, "A deep learning system for sentiment analysis of service calls," *arXiv preprint arXiv:2004.10320*, 2020.

[26] S. A. Vermeer, T. Araujo, S. F. Bernritter, and G. van Noort, "Seeing the wood for the trees: How machine learning can help firms in identifying relevant electronic word-of-mouth in social media," *International Journal of Research in Marketing*, vol. 36, no. 3, pp. 492–508, 2019.

[27] "'ask jamie' virtual assistant," Mar 2022. [Online]. Available: https://www.tech.gov.sg/products-and-services/ask-jamie/

[28] A. I. Niculescu, I. Kukanov, and B. Wadhwa, "Digimo-towards developing an emotional intelligent chatbot in singapore," in *Proceedings of the 2020 Symposium on Emerging Research from Asia and on Asian Contexts and Cultures*, 2020, pp. 29–32.

[29] J. Lin, S. Chinthavali, C. D. Stahl, C. Stahl, S. Lee, and M. Shankar, "Ecosystem discovery: Measuring clean energy innovation ecosystems through knowledge discovery and mapping techniques," *The Electricity Journal*, vol. 29, no. 8, pp. 64–75, 2016.

[30] J. Xiong, Y. Hswen, and J. A. Naslund, "Digital surveillance for monitoring environmental health threats: A case study capturing public opinion from twitter about the 2019 chennai water crisis," *International journal of environmental research and public health*, vol. 17, no. 14, p. 5077, 2020.

[31] M. Alam, F. Abid, C. Guangpei, and L. Yunrong, "Social media sentiment analysis through parallel dilated convolutional neural network for smart city applications," *Computer Communications*, vol. 154, pp. 129–137, 2020.

[32] D. Pérez-Fernández, J. Arenas-García, D. Samy, A. Padilla-Soler, and V. Gómez-Verdejo, "Corpus viewer: Nlp and ml-based platform for public policy making and implementation," 2019.

[33] R. Bhatt, M. Patel, G. Srivastava, and V. Mago, "A graph based approach to automate essay evaluation," in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2020, pp. 4379–4385.

[34] A. Miaschi, D. Brunato, and F. Dell'Orletta, "A nlp-based stylometric approach for tracking the evolution of l1 written language competence." *Journal of Writing Research*, vol. 13, no. 1, 2021.

[35] N. Goenawan and C. Wong, "Duolingo shared task on second language acquisition modeling (slam)."

[36] S. Mayhew, K. Bicknell, C. Brust, B. McDowell, W. Monroe, and B. Settles, "Simultaneous translation and paraphrase for language education," in *Proceedings of the fourth workshop on neural generation and translation*, 2020, pp. 232–243.

[37] L. Alrajhi, K. Alharbi, and A. I. Cristea, "A multidimensional deep learner model of urgent instructor intervention need in mooc forum posts," in *International conference on intelligent tutoring systems*. Springer, 2020, pp. 226–236.

[38] A. Poce, F. Amenduni, M. R. Re, C. De Medio, and A. Norgini, "Correlations among natural language processing indicators and critical thinking sub-dimensions in hied students," *Form@ re-Open Journal per la formazione in rete*, vol. 20, no. 3, pp. 43–67, 2020.

[39] I. I. Bogoch, O. J. Brady, M. U. Kraemer, M. German, M. I. Creatore, M. A. Kulkarni, J. S. Brownstein, S. R. Mekaru, S. I. Hay, E. Groot *et al.*, "Anticipating the international spread of zika virus from brazil," *The Lancet*, vol. 387, no. 10016, pp. 335–336, 2016.

[40] S. Enarvi, M. Amoia, M. D.-A. Teba, B. Delaney, F. Diehl, S. Hahn, K. Harris, L. McGrath, Y. Pan, J. Pinto *et al.*, "Generating medical reports from patient-doctor conversations using sequence-to-sequence models," in *Proceedings of the first workshop on natural language processing for medical conversations*, 2020, pp. 22–30.

[41] R. A. Li, I. Hajjar, F. Goldstein, and J. D. Choi, "Analysis of hierarchical multi-content text classification model on b-sharp dataset for early detection of alzheimer's disease," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 2020, pp. 358–365.

[42] H. Öztürk, A. Özgür, P. Schwaller, T. Laino, and E. Ozkirimli, "Exploring chemical space using natural language processing methodologies for drug discovery," *Drug Discovery Today*, vol. 25, no. 4, pp. 689–705, 2020.

[43] B. Inkster, S. Sarda, V. Subramanian *et al.*, "An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: real-world data evaluation mixed-methods study," *JMIR mHealth and uHealth*, vol. 6, no. 11, p. e12106, 2018.

[44] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial," *JMIR mental health*, vol. 4, no. 2, p. e7785, 2017.

[45] R. Meadows, C. Hine, and E. Suddaby, "Conversational agents and the making of mental health recovery," *Digital health*, vol. 6, p. 2055207620966170, 2020.

[46] L. Yang, J. Zeng, T. Peng, X. Luo, J. Zhang, and H. Lin, "Leniency to those who confess? predicting the legal judgement via multi-modal analysis," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, 2020, pp. 645–649.

[47] F. Ruggeri, F. Lagioia, M. Lippi, and P. Torroni, "Detecting and explaining unfairness in consumer contracts through memory networks," *Artificial Intelligence and Law*, vol. 30, no. 1, pp. 59–92, 2022.

[48] M. Queudot, É. Charton, and M.-J. Meurs, "Improving access to justice with legal chatbots," *Stats*, vol. 3, no. 3, pp. 356–375, 2020.

[49] K. Sugathadasa, B. Ayesha, N. d. Silva, A. S. Perera, V. Jayawardana, D. Lakmal, and M. Perera, "Legal document retrieval using document vector embeddings and deep learning," in *Science and information conference*. Springer, 2018, pp. 160–175.

[50] D. Jain, M. D. Borah, and A. Biswas, "Fine-tuning textrank for legal document summarization: A bayesian optimization based approach," in *Forum for Information Retrieval Evaluation*, 2020, pp. 41–48.

[51] A. Alshehri, E. M. B. Nagoudi, and M. Abdul-Mageed, "Understanding and detecting dangerous speech in social media," *arXiv preprint arXiv:2005.06608*, 2020.

[52] O. Araque and C. A. Iglesias, "An ensemble method for radicalization and hate speech detection online empowered by sentic computing," *Cognitive Computation*, vol. 14, no. 1, pp. 48–61, 2022.

[53] I. Percy, A. Balinsky, H. Balinsky, and S. Simske, "Text mining and recommender systems for predictive policing," in *Proceedings of the ACM Symposium on Document Engineering 2018*, 2018, pp. 1–4.

[54] D. Sun, X. Zhang, K.-K. R. Choo, L. Hu, and F. Wang, "Nlp-based digital forensic investigation platform for online communications," *Computers & Security*, vol. 104, p. 102210, 2021.

[55] C. Ziems, Y. Vigfusson, and F. Morstatter, "Aggressive, repetitive, intentional, visible, and imbalanced: Refining representations for cyberbullying classification," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, 2020, pp. 808–819.

[56] A. Daly, T. Hagendorff, H. Li, M. Mann, V. Marda, B. Wagner, W. W. Wang, and S. Witteborn, "Artificial intelligence, governance and ethics: global perspectives," *The Chinese university of Hong Kong faculty of law research paper*, no. 2019-15, 2019.

[57] M. Jain, P. Kumar, R. Kota, and S. N. Patel, "Evaluating and informing the design of chatbots," in *Proceedings of the 2018 Designing Interactive Systems Conference*, 2018, pp. 895–906.

[58] J. Weizenbaum, "Eliza—a computer program for the study of natural language communication between man and machine," *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966.

[59] T. Brennan, W. Dieterich, and W. Oliver, "Compas: Correctional offender management for alternative sanctioning," *Technical manual and psychometric report*, vol. 5, 2007.

Satyam Mohla received the B.Tech & M.Tech degree in Electrical Engineering from Indian Institute of Technology, Bombay. India. He is currently affiliated with Value Creation Division, Digital Transformation Supervisory Unit, Honda Innovation Lab Tokyo specialising in autonomous vehicles, digital transformation, philosophy and business. He was a Salzburg Global Fellow, a Shastri Fellow at Shastri Indo-Canadian Institute and Temasek TfLEaRN Scholar at NTU Singapore.

Anupam Guha received the PhD in Computer Science from University of Maryland in 2017, & MS in Computer Science from Georgia Tech in 2010. He is an Assistant Professor with the Centre for Policy Studies, IIT Bombay. He specialises in working at the intersection of language and vision, and how AI systems work and fail around these problems.