Regulation and NLP (RegNLP): Taming Large Language Models

Catalina Goanta
Utrecht University

Nikolaos Aletras
University of Sheffield

Ilias Chalkidis
University of Copenhagen

Sofia Ranchordas
Tilburg University

Gerasimos Spanakis
Maastricht University

Abstract

The scientific innovation in Natural Language Processing (NLP) and more broadly in artificial intelligence (AI) is at its fastest pace to date. As large language models (LLMs) unleash a new era of automation, important debates emerge regarding the benefits and risks of their development, deployment and use. Currently, these debates have been dominated by often polarized narratives mainly led by the *AI Safety* and *AI Ethics* movements. This polarization, often amplified by social media, is swaying political agendas on AI regulation and governance and posing issues of regulatory capture. Capture occurs when the regulator advances the interests of the industry it is supposed to regulate, or of special interest groups rather than pursuing the general public interest. Meanwhile in NLP research, attention has been increasingly paid to the discussion of regulating risks and harms. This often happens without systematic methodologies or sufficient rooting in the disciplines that inspire an extended scope of NLP research, jeopardizing the scientific integrity of these endeavors. *Regulation studies* are a rich source of knowledge on how to systematically deal with *risk and uncertainty*, as well as with *scientific evidence*, to evaluate and compare regulatory options. This resource has largely remained untapped so far. In this paper, we argue how NLP research on these topics can benefit from proximity to regulatory studies and adjacent fields. We do so by discussing basic tenets of regulation, and risk and uncertainty, and by highlighting the shortcomings of current NLP discussions dealing with risk assessment. Finally, we advocate for the development of a new multidisciplinary research space on regulation and NLP (RegNLP), focused on connecting scientific knowledge to regulatory processes based on systematic methodologies.

1 Introduction

The development of Large Language Models (LLMs) is at its fastest pace to date. In the past years alone, LLMs have seen considerable advancement across a multitude of languages and types of data, with models such as GPT-3.5 (Ouyang et al., 2022), GPT-4 (OpenAI, 2023), LLaMA (Touvron et al., 2023), and PALM-2 (Anil et al., 2023) demonstrating unprecedented capabilities across a broad collection of natural language processing (NLP) tasks.[1]

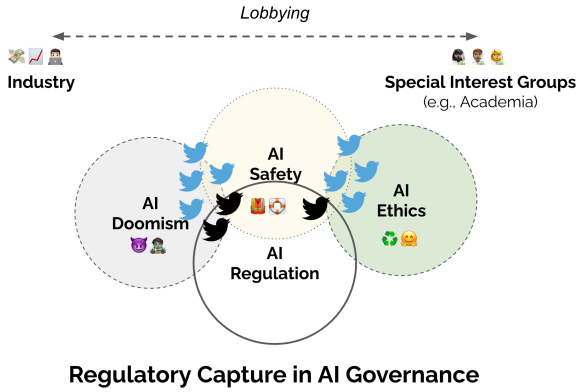

Figure 1: A depiction of the cross-over between AI Safety, Ethics, Doomism, and how they capture AI Regulation.

These innovations have led to rapid shifts in various applications such as open-domain search, coding, e-commerce and education. For example, state-of-the-art LLMs already power conversational search engines (e.g. OpenAI ChatGPT, Bing Chat, and Google Bard), coding assistants (e.g. OpenAI Codex and Github Copilot), product recommender systems (e.g. Alibaba Tongyi and SalesForce CommerceGPT) and educational assistants (e.g. Khanmigo) inter alia.

As with most technologies, the development and use of LLMs do not come without concerns. Researchers are rightfully worried that while this technology may be transformative, its societal implications might be higher than its benefits (Gabriel, 2020). Concerns have been raised es-

[1] In the sense that LLMs generalize to a great extend in out-of-distribution and out-of-domain use cases.

pecially around ethics (Floridi, 2023; Tsarapatsanis and Aletras, 2021), bias (Hovy and Prabhumoye, 2021; Blodgett et al., 2020), safety (Dobbe et al., 2021) and environmental impact (Rillig et al., 2023; Schwartz et al., 2020; Strubell et al., 2019). The unsupervised use of LLMs has already led to widely-publicized examples of professional negligence. The all too public fiasco of the lawyer who used ChatGPT for a court brief and unknowingly included made-up case law references was taken by many as an example of the dangers of immersing daily professional activities in generative AI.[2] The public debate around the future of AI, and implicitly NLP, is very complex and multi-layered. While the debate seems to converge on the point of calling for regulation to control the unwanted effects of these technologies,[3] different regulatory directions are proposed by the various stakeholders involved in this debate.

The current public discourse has been dominated by two groups. On the one hand, proponents of AI existential risks (the *AI Safety* movement)[4] that include technology CEOs and AI researchers have been publishing open letters[5,6,7] and regularly meet with regulators to warn about catastrophic scenarios around general AI, proposing industry-friendly solutions (Roose, 2023). On the other hand, the *AI ethics* movement (Borenstein et al., 2021), mostly reflects the voice of researchers from various disciplines as well as civil society activists. They have raised compelling alarm bells with respect to the risks posed by LLMs (Bender et al., 2021; Weidinger et al., 2022; Floridi, 2023). The AI ethics movement has also offered guidance to regulators such as the European Parliament on AI governance, arguing for e.g. broad definitions of general purpose AI.[8]

Yet while NLP research increasingly focuses on LLM regulation, it remains generally detached from prior work on regulation studies. Instead, the often intense public conflicts in this space are nudging regulators towards reactionary public relations activities rather than the collection of scientific expertise representing broader parts of the NLP and AI communities. For instance, right after the most recent letter on existential risks, the US and EU agreed to develop an AI code of conduct 'within weeks'.[9] Similarly, the UK has announced it will be holding a global summit on AI safety,[10] and the US Congress has been taking evidence from a wide array of industry executives, in what has been described as a global race to regulate AI. [11] With legacy and social media considerably increasing the visibility of these often polarizing debates, there is a real danger of regulatory capture by visible voices in industry and academia alike on scientific views that might not necessarily be representative of the 'silent majority' of NLP researchers. Regulatory capture is the process by which regulation is directed away from the public interest and towards the interests of specific groups (Levine and Forrence, 1990; Dal Bo, 2006).

Against this background, we call on the NLP community to familiarize itself with regulation studies. We argue that this can lead to a clearer vision about how NLP as a field can properly participate in AI governance not only as an object of regulation, but also as a source of scientific knowledge that can benefit individuals, societies and markets alike. In this paper, we contribute to the existing debate relating to the future of NLP by discussing the benefits of interfacing NLP research with regulation studies in a systematic way. This view is based on two main ideas:

1. NLP research on regulation needs a multidisciplinary framework engaging with regulation studies, as well as adjacent disciplines such as law, economics, environmental science, etc. We advocate for a new crucial area of research on regulation and NLP (RegNLP), with harmonized and systematic methodologies.

2. A more coordinated NLP research field on risk

[2] https://edition.cnn.com/2023/05/27/business/chat-gpt-avianca-mata-lawyers/index.html

[3] https://www.forbrukerradet.no/side/new-report-generative-ai-threatens-consumer-rights/

[4] We distinguish the AI Safety movement from 'AI Doomism' (https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/) flirting with conspiracy theories.

[5] https://www.nytimes.com/2023/03/29/technology/ai-artificial-intelligence-musk-risks.html

[6] https://futureoflife.org/open-letter/pause-giant-ai-experiments/

[7] https://www.safe.ai/statement-on-ai-risk

[8] https://ainowinstitute.org/publication/gpai-is-high-risk-should-not-be-excluded-from-eu-ai-act.

[9] https://www.fastcompany.com/90903919/will-the-eu-u-s-new-voluntary-code-of-conduct-on-ai-work-to-rein-in-the-tech

[10] https://www.theguardian.com/technology/2023/jun/09/rishi-sunak-ai-summit-what-is-its-aim-and-is-it-really-necessary.

[11] https://foreignpolicy.com/2023/05/05/eu-ai-act-us-china-regulation-artificial-intelligence-chatgpt/

and regulation (such as RegNLP) can interact with policy-makers with more transparency, representation, and trustworthiness.

2 Regulation: A Short Introduction

Why Do We Regulate? Calls to regulate LLMs and AI are everywhere, to the extent that overusing the term 'regulation' is trivializing its meaning. So what exactly *is* regulation and why do we rely on it?

Historically, regulation was defined by reference to state intervention in the economy. Selznick (1985) defined regulation as 'a sustained and focused control exercised by a public agency over activities that are valued by the community'. Over the last decades, regulation has evolved and it has increasingly acquired a hybrid character as both public and private actors may issue rules that shape social behavior. In this paper, we draw on Black's definition of regulation as an organized and intentional attempt to manage another person's behavior so as to solve a collective problem (Black, 2008). This is done through a combination of rules or norms which come together with means for their implementation and enforcement (Ogus, 2009).

Regulation includes different types of regulatory instruments (e.g. laws) such as traditional top-down or command-and-control regulations. Recent examples in digital public policy include the laws issued by the European Union, such as the Digital Services Act, or India's law banning TikTok. However, beyond these public regulatory instruments, there is an array of private or hybrid instruments (Veale et al., 2023). Some of these are qualified as 'soft' regulation because they cannot be enforced in court but they remain relevant and they effectively shape the behavior of the industry. Examples include the EU's Ethics Guidelines for Trustworthy AI[12], but also codes of conduct initiated by industry itself.[13]

Over the last decades, different theories of regulation have helped us understand either on normative or empirical accounts why we should regulate. They generally reflect various hypotheses about 'why regulation emerges, which actors contribute to that emergence and typical patterns of interaction between regulatory actors' (Morgan and Yeung, 2007). Public interest and private interest theories are two well-known examples (Morgan and Yeung, 2007). According to public interest theories, regulation is used by law-makers to serve a broad public interest, seeking to regulate in the most efficient way possible. Regulators assume that markets need 'a helping hand' because when unhindered, they will fail. Information asymmetries are one of the market failures that regulation seeks to address. At the same time, public interest theories of regulation are normative and prescriptive: on the one hand, they assume that benevolent state regulators ought always to use regulation to advance the public interest; on the other, they also advise on how to achieve this goal.

In contrast, private interest theories are not primarily concerned with normative justifications for regulation. Rather, they are prescriptive accounts of the complex dynamics between different market actors, stakeholders and public officials situated in a given socio-economic, political and cultural time and place. They explain, for example, why regulation may fail to pursue the public interest but they do not offer prescriptions on how address to such problems. Private interest theories assume that regulation emerges from the actions of individuals or groups motivated to maximize their self-interest.

Technology Regulation and the Role of Science Technological change often disrupts the wider regulatory order, triggering concerns about its adequacy and regulatory legitimacy (Brownsword et al., 2017). Differences in the timing of technology and regulation explain this difficulty. The literature has claimed there is a 'pacing gap' between the slow-going nature of regulation and the speed of technological change (Marchant et al., 2013). Technological innovations have specific development trajectories, investment and life cycles, and path dependencies (van den Hoven, 2014) that do not go well with the speed of technology. This also applies to LLMs. This is a well-known problem in regulatory studies that has been captured by the Collingridge dilemma (Genus and Stirling, 2018). This dilemma explains that when an innovation emerges, regulators hesitate to regulate due to the limited availability of information. However, by the time more is known, regulations may have become obsolete as technology may have already changed.

The increased pace of regulatory activities in the past years in the field of technology shows that regulators are trying to close the pacing gap and

[12] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[13] https://ethics.acm.org/code-of-ethics/software-engineering-code/

be more proactive in tackling the potential risks of technology. In doing so, they increasingly depend on retrieving scientific information in a quick and agile way.

The role of science in regulation and public policy has been the subject of important debates. On the one hand, regulation should reflect the latest scientific evidence and be evidence-based. An illustration of this approach is the European Union's 'Better Regulation' agenda, a public policy strategy aiming to ensure that European regulation is based on scientific evidence, as well as the involvement of a wide range of stakeholders in the decision-making process (Commission, 2023b; Simonelli and Iacob, 2021). Citizens, businesses and any other stakeholders can submit their contributions to the calls for evidence, feedback and public consultations. For instance, 303 contributions were received on the AI Act proposal during 26 April 2021 and 6 August 2021, out of which 28% from businesses, 24% from business associations, 17% from NGOs, and 6% from academic/research institutions (Commission, 2023a). On the other hand, science is complex and difficult to translate into regulatory measures. Science has thus been used to oversimplify regulatory problems and justify poor regulatory decisions based on the existence of scientific evidence pointing in a specific direction (Porter, 2020). This is a particular danger in the LLM public debate. With science becoming increasingly complex, so do the scientific perspectives on how to proceed with this technology.

Lessons to be Learned The theoretical underpinnings of regulation have helped shape a cohesive understanding of the rationale behind regulatory activity. The interest in technology regulation, particularly for disruptive innovations such as LLMs, has exploded in past years across a wide array of scientific disciplines. While understandable, such popularity often leads to inquiries which are not connected to prior knowledge on regulation studies. This reflects a more general problem faced by contemporary science, namely that of tackling multidisciplinary issues without multidisciplinary expertise. Considering these circumstances, research on LLMs and regulation could benefit from engaging with the regulation studies and governance context.

3 LLMs: Risk and Uncertainty

The need to bridge NLP research with regulation studies is especially important in the discussion of risks. New and emerging technologies are typically accompanied by risk and uncertainty. The regulation of technological change and innovation is highly complex, as innovation remains an elusive concept hard to define, measure, and thus regulate.

In the past years, the question of risks arising out of NLP developments such as LLMs has been increasingly embraced in computer science literature. One strand of this literature is reflected by the theme of algorithmic unfairness. This theme emerged at the intersection of discrimination law and automated decision-making, and includes questions relating to fairness and machine learning in general (Barocas et al., 2019), as well as specific examples of algorithmic bias risks in NLP (Field et al., 2023; Talat et al., 2022; Kidd and Birhane, 2023), computer vision (Wolfe et al., 2023), multimodal models (Birhane et al., 2021), as well as privacy risks (Mireshghallah et al., 2022). Another strand of this literature looks at LLMs from a more holistic perspective, raising concerns about their size vis-a-vis a broader number of risks for e.g. the environment, bias, representation or hate speech (Bender et al., 2021; Weidinger et al., 2022; Bommasani et al., 2022). This theme does not only include risks in commercial applications, but also risks arising out of the mere scientific development of technology.

This literature has raised important concerns relating to the immediate and longer-term implications around the advancement of machine learning and NLP. However, when positioned in the regulatory context, we can observe conceptual clashes with frameworks which have been traditionally relied upon in public policy and risk regulation. One such framework is the field of risk and uncertainty. Put simply, 'risk is the situation under which the decision outcomes and their probabilities of occurrences are known to the decision-maker, and uncertainty is the situation under which such information is not available to the decision-maker' (Park and Shapira, 2017). In more technical terms, 'risk is the probability of an event multiplied by its impact, and uncertainty reflects the accuracy with which a risk can be assessed' (Krebs, 2011). As a field, risk and uncertainty has made considerable contributions to the development of risk regulation, most notably in relation to environmental regula-

tion (Heyvaert, 2011; Yarnold et al., 2022). It is important that policy-makers have a concrete quantification of risk (Aumann and Serrano, 2008), in order to determine the adequate level of risk associated with various public policies. In addition, risk determination and management have important economic consequences, specifically for determining 'what level of expenditure in reducing risk is proportionate to the risk itself' (Krebs, 2011).

Especially in the case of technologies that easily transcend physical borders, societies and economies, determining risk and uncertainty is a complex undertaking, even for scientists. Some of the factors that make it difficult to assess risk and uncertainty include the complexity of the technology itself, as well as the information asymmetry underlying commercial practices. LLMs are humongous (billion-parameters-sized) Transformer-based (Vaswani et al., 2017) models, which have been initially pre-trained as standard language models (Radford et al., 2018) in a vast quantity of text data (mostly web scrapes) and have been also further optimized to follow instructions (Chung et al., 2022) and user *alignment* (Leike et al., 2018) with reinforcement learning from human feedback (RLHF) (Christiano et al., 2017; Stiennon et al., 2020). Particularly, AI alignment has been a controversial topic, since it implies a broad consensus on what sort of values (standards) AI should align with (Gabriel, 2020). As such, LLMs are complex technologies where in addition to risk, we also deal with considerable uncertainty, which can be, among others, descriptive (e.g. relating to the variables defining a system), or related to measurement (e.g. uncertainty about the value of the variables in a system) (Gough, 1988).

LLMs as the new GMOs? Yet LLMs are neither the first nor the last technology development posing concerns about wide-spread risks. As an illustration, in the 1970s and 1980s, In-Vitro Fertilization (IVF) was a demonized scientific development considered inhumane, which led to nothing short of a large-scale moral panic (Garber, 2012). In the 2000s, concerns around Genetically Modified Organisms (GMOs) dominated media coverage in European and North American countries, in what was deemed a 'superstorm' of moral panic and new risk discourses (Howarth, 2013). These are only two examples of risk narratives that were amplified by media coverage in ways that overshadowed important scientific expertise. Yet through regulation supporting scientific advancement, their use today has become mundane as part of solving considerable society problems such as infertility or food availability. These comparisons by no means imply that earlier biotechnological innovations pose the same levels of risk as LLMs or should entail the same level of regulation. However, it is important to learn from our past experiences with technology how to distinguish between moral panics and real problems that need scientific solutions.

Researchers are starting to develop concrete methodologies for the auditing of LLMs and related NLP technologies (Derczynski et al., 2023),[14] as well as dealing with particular risks such as environmental impact (Rolnick et al., 2022). These contributions are much needed, as they can be translated into concrete measurements of risk and uncertainty, and further lead to the development of policy options in risk management. However, these initiatives are so far too few, as no cohesive scientific approach exists on the assessment of the risk and uncertainty posed by LLMs. To date, even the most comprehensive overviews of LLM risks (Weidinger et al., 2022) lack basic methodological practices such as the systematic retrieval of information from the disciplines of inquiry (Page et al., 2021). In some cases, strong projections about risk impact are made without any scientific rigor whatsoever (Hendrycks et al., 2023). Similarly, Dobbe et al. (2021) note that while many technical approaches, including approaches related to 'mathematical criteria for "safety" or "fairness" have started to emerge, 'their systematic organization and prioritization remains unclear and contested.

As a result, the lack of systematization and methodological integrity in scientific work around LLM risks contributes to a credibility crisis which may impact regulation and governance directly.

Innovation Governance and Risk Relativization Learning from earlier experiences with risk and uncertainty can also help NLP researchers understand how risk has been dealt with in other policy areas. One of the reasons why it is important to contextualize NLP research on risks into a broader regulatory landscape is because this area has already generated meaningful frameworks for the understanding of risk in the context of innovation governance. Such frameworks include for instance the principle of responsible innovation, which calls for 'taking care of the future through collective steward-

[14] https://github.com/leondz/garak/

ship of science and innovation in the present' (von Schomberg, 2013). This is only one of the many other approaches that can guide decision-making on technology regulation (Hemphill, 2020).

A regulatory angle can also help with the relativization of risk - that is, putting risks into perspective by considering other policy areas as well. For instance, the UK Risk Register 2020 discusses potential risks and challenges that could cause significant disruption to the UK [15]. The report thematically groups six risks (malicious attacks, serious and organized crime, environmental hazards, human and animal health, major accidents and societal risks). This insight is useful in understanding the scale and diversity of risks that public policy needs to account for. Such awareness could also contribute to the generation of policy options that can put LLM risks into perspective in relation to other categories of risks like the ones mentioned above. Here, a more holistic perspective on risk could also take a sectoral approach to LLM risks. For instance, going back to the example of the lawyer who invented case law using ChatGPT, existing legal and self-regulatory frameworks already address the risk of negligence in conducting professional legal activities. Considering this context can contribute with insights into whether it is really necessary to treat LLM-mediated information as a novel danger. While technology has generated a broad digital transformation (Verhoef et al., 2021), it adds layers to existing problems (e.g. social inequality) which need policy interventions independent from their digital amplification.

4 Scientific Expertise, Social Media and Regulatory Capture

Regulatory Capture As a source of evidence for policy-makers, scientific expertise has increasingly played a central role in regulation (Paschke et al., 2019). In some supranational governance contexts, scientific expertise is called upon in procedures that often require a certain level of transparency. This is the case of the call for public comments which we have discussed above as part of the EU's 'Better Regulation' agenda. However, in the past decade, the rise of science communication on social media has somewhat changed the interaction between policy-makers and scientists (van Dijck and Alinejad, 2020). A lot of the public debate between stakeholders from industry, academia and policy relating to LLM risks is had on social media platforms such as Twitter. This can pose a regulatory capture problem. Regulatory capture occurs when the regulator advances the interests of the industry it is supposed to regulate, or of special interest groups rather than pursuing the public interest (Carpenter and Moss, 2013). Regulatory capture fits within the private interest theories we explored in Section 2, and often refers to the influence exercised by industry over regulation processes (Saltelli et al., 2022). The most recent example is OpenAI's white paper suggesting narrow regulatory interpretations for general purpose high-risk AI systems to European regulators [16]. Similarly, multiple technology executives of large companies using LLMs, such as HuggingFace or OpenAI have been testifying before the US Congress to propose industry-friendly interpretations of AI risks. This is happening in a context of existing concerns around the industry orchestration of research agendas in NLP (Abdalla et al., 2023), and science (Abdalla and Abdalla, 2021) in general. However, regulation can also be captured by special interest groups from civil society, and increasingly, academia. In this meaning, regulatory capture has a cultural or value-driven dimension that encompasses 'intellectual, ideological, or political forms of dominance' (Saltelli et al., 2022). In a landscape where special interest groups are increasingly represented by popular science communicators, a lot of questions arise in relation to the power exercised by the rising impact of science influencers, whether from academic, journalism and industry environments (Zhang and Lu, 2023).

The Rise of Science Influencers Traditionally, science communication has followed a conventional model dominated by professional actors gatekeeping information (e.g. scientists, journalists and government). Social media has led to the creation of a networked model of science communication, with underlying socio-technical and political power shifts (van Dijck and Alinejad, 2020). Science influencers are rooted in this development, as well as the broader rise of social media influencers (Goanta and Ranchordás, 2020). They also bring with them additional complexities. They may emerge from a scientific background, but may use their platforms both for professional as well as personal

[15] https://www.gov.uk/government/publications/national-risk-register-2020

[16] https://time.com/6288245/openai-eu-lobbying-ai-act/

self-disclosure (Kim and Song, 2016). In doing so, they also become political influencers who 'harness their digital clout to promote political causes and social issues' (Riedl et al., 2023). To signal the social media influence of such public opinion leaders, some media outlets even rank them for awards purposes,[17] or profile them vis-a-vis state of the art scientific expertise.[18]

This development is especially important since some social media platforms are more relevant for regulators than others. A recent report published by Oxford University's Reuters Institute shows that Twitter is the platform politicians pay most attention to across all studied markets.[19] With this in mind, the polarized narratives around LLM risks unfolding on social media pose the danger that scientific expertise is only partially represented in public debates, in spite of the promises of speech democratization expected from these platforms in the context of science communication.

While followers and engagement may be a measure of popularity with some communities and stakeholders, it raises concerns relating to the role popularity metrics and algorithmic amplification on social media may have in representing scientific or industry consensus before policy-makers. As Zhang et al. (2018) put it, there is a popularity bias that means 'attention tends to beget attention'. In other words, 'the more contacts you have and make, the more valuable you become, because people think you are popular and hence want to connect with you' (van Dijck, 2013).

How exactly scientific popularity influences regulation still needs to be explored in greater detail, particularly as a novel example of potential regulatory capture. What we know so far is that social media influencers can be highly effective in relying on authenticity and para-social relations for persuasion purposes (Vannini and Franzese, 2008; Hudders et al., 2021). In this context, popularity determines power relationships within social media networks that may capture regulatory processes in two ways. First, by exercising persuasion over policy-makers as audiences through visibility and popularity. Simply put, not all research that is available in a given field is presented on social media. Under the premise of basing policy on scientific evidence, politicians may rely on research that gains visibility due to amplification by scientific influencers, particularly when social media popularity is doubled by the brand power of prestigious academic institutions. Second, by amplifying polarized debates that may trigger policy options which are not sufficiently informed through transparent and collective processes of evidence gathering. If multiple AI research groups are vocal on social media about the future of AI research, this fuels a race towards AI regulation. This can take away from the thoroughness that is necessary in collecting evidence for such a complex field.

5 Regulation and NLP (RegNLP): A New Field

The danger of regulatory capture, taken together with the lack of systematization in the identification and measurement of risk and uncertainty around LLMs, calls for a cohesive scientific agenda and strategy. In 2023, the world counts 8 billion humans[20], and further digitalization and automation are not only unstoppable, but also absolutely necessary. Technological innovation is currently vital in the governance of our society. That does not mean its pursuit ought to be free from regulatory frameworks mandating rules on how to deal with the risks it poses. NLP research has already drawn attention to some of the potential risks of LLMs. To consolidate this effort, it is necessary to consider in what direction NLP research can further develop and what contributions it can make to regulation. A concrete proposal we advocate for is the creation of a new field of scientific inquiry which we call Regulation and NLP (RegNLP). RegNLP has three essential features which we discuss below.

5.1 Multidisciplinarity

First, RegNLP needs to be a multidisciplinary field that spans across any scientific areas of study which are relevant for the intersection of regulation and AI. In the past years, multidisciplinary communities have been increasingly popular. An example is the ACM Fairness, Accountability and Transparency Conference (FAccT[21]), which often features NLP research. Such research communities

[17] https://www.euractiv.com/section/digital/news/meet-the-2019-euinfluencer-awardees/

[18] https://spectrum.ieee.org/artificial-general-intelligence

[19] https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2023

[20] https://ourworldindata.org/world-population-growth. For a brief comparison, at the time of the Dartmouth AI workshop in 1956, the world population was at a mere 2.5 billion, Statista, 2023.

[21] https://facctconference.org

form around the disciplines that are most reflected by their research questions. For RegNLP, the constitutive disciplines ought to include NLP and regulation studies but also law, economics, political science, etc.. NLP research approaches cannot replace expertise from other fields. At the same time, expertise entails more than an interest in an adjacent field, but rather a deep understanding of the contributions and limitations such a field can entail when interacting with NLP. One strategy to encourage this cross-pollination is for NLP researchers interested in regulation to co-author papers with regulation experts and other relevant scholars. In doing so, RegNLP can also contribute to the research gaps in other fields, such as public administration, where literature on AI still needs further development. For instance, in 2019, only 12 scientific articles were published on AI and public policy and administration, mostly focused on the use of AI *in* public administration (Valle-Cruz et al., 2020). Similarly, a quick search in 'Regulation & Governance', a leading journal in regulation studies, yields a total of 18 results, out of which only one discusses AI risks (Laux et al., 2023).

5.2 Harmonized Methodologies

RegNLP needs harmonized methodologies. One of the biggest problems with the consolidation of multidisciplinary research agendas and communities relates to the lack of alignment between the different methods and goals pursued by different researchers. This issue trickles down into all relevant activities which normally help consolidate multidisciplinary groups, and is most specifically visible in the process of peer review. If reviewers are not familiar with methodologies from other fields, they will be unable to adequately assess the quality of research (Laufer et al., 2022). This can lead to the publication of research which may be interesting across disciplines, but which may not meet the methodological rigor of the scientific discipline to which a given method pertains.

RegNLP can help establish shared standards for scientific quality around shared methodologies and science practices. This can mean embracing a diverse methodological scope to reflect the tools that are needed in the inquiry of different types of research questions. It can also mean perfecting existing methods and deploying them on novel sources of data, as NLP research methods are a natural starting point for the systemic retrieval of complex information and overviews from existing scientific research, such as meta-studies (Heijden, 2021).

5.3 Science Participation in Regulation

RegNLP can help research on regulation and NLP interface with regulatory processes. At a time of increased complexity, it is important for scientists to clarify the state of art of fast paced technological change. Using harmonized methodologies in the context of a multidisciplinary research agenda can bring much needed coordination to the interaction between NLP development and regulation. In the absence of such coordination, as we have discussed in Section 4, there is a potential danger that policy-makers are only exposed to popular scientific opinions instead of consolidated science communication. What is more, embedding RegNLP into a risk and regulation context can offer further inspiration for the role of academia, as a repository of public trust. A lot of regulatory agencies and standardization bodies govern the implementation of regulation. In addition, new forms of interactions with civil society are being set up by EU regulation, such as the Digital Services Act's 'trusted flaggers', namely organizations that can flag illegal content on online platforms. Similarly, there can be new roles to play for RegNLP agendas and communities.

6 Conclusion

LLMs reflect a momentous development in NLP research. As they unleash a new era of automation, it is important to understand their risks and how these risks can be controlled. While eager to engage with regulatory matters, NLP research on LLM risks has so far been disjointed from other fields which are of direct interest, such as regulation studies. In particular, the field of risk and uncertainty has been conceptualizing and discussing scientific risks for decades. In this paper, we introduced these two areas of study and explained why it would be beneficial for NLP research to consider them in greater depth. In doing so, we also raised concerns relating to the fact that a lot of scientific debates on NLP risks are taking place on social media. This may lead to regulatory capture, or in other words the exercise of influence over law-makers, who are notoriously active on social media platforms. To tackle these issues, we propose a new multidisciplinary area of scientific inquiry at the intersection of regulation and NLP (RegNLP), aimed at the

development of a systematic approach for the identification and measurement of risks arising out of LLMs and NLP technology more broadly.

Limitations

Our paper reflects on the future of NLP in a landscape where interest in regulation is increasing exponentially, within and outside the field. Given the nature of this paper, we will refer to limitations dealing with the feasibility of our proposed research agenda. The most important limitation reflects the discussion around inter- and multidisciplinarity. This is by no means a new theme in science, and its implementation has cultural, managerial and economic implications that we do not discuss in the paper, but which are important to acknowledge. Similarly, another limitation is reflected by the modest amount of knowledge we have relating to the impact of social media influencers (such as science influencers) on regulation and public policy. In this paper, we raise certain issues around science communication as the starting point of a broader discussion around power and influence in law-making as amplified by social media.

Acknowledgements

Ilias Chalkidis is funded by the Novo Nordisk Foundation (grant NNF 20SA0066568). Sofia Ranchordas is funded by the NWO-Vidi project 'Vulnerability in the Digital Administrative State'. Catalina Goanta is funded by the ERC Starting Grant HUMANads (ERC-2021-StG No 101041824).

References

Mohamed Abdalla and Moustafa Abdalla. 2021. The grey hoodie project: Big tobacco, big tech, and the threat on academic integrity. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 287–297, New York, NY, USA. Association for Computing Machinery.

Mohamed Abdalla, Jan Philip Wahle, Terry Ruas, Aurélie Névéol, Fanny Ducel, Saif M. Mohammad, and Karën Fort. 2023. The elephant in the room: Analyzing the presence of big tech in natural language processing research.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Robert J Aumann and Roberto Serrano. 2008. An economic index of riskiness. *J. Polit. Econ.*, 116(5):810–836.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org. http://www.fairmlbook.org.

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

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes.

Julia Black. 2008. Constructing and contesting legitimacy and accountability in polycentric regulatory regimes. *Regul. Gov.*, 2(2):137–164.

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

5476, Online. Association for Computational Linguistics.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the opportunities and risks of foundation models.

Jason Borenstein, Frances S Grodzinsky, Ayanna Howard, Keith W Miller, and Marty J Wolf. 2021. AI ethics: A long history and a recent burst of attention. *Computer (Long Beach Calif.)*, 54(1):96–102.

Roger Brownsword, Eloise Scotford, and Karen Yeung, editors. 2017. *The oxford handbook of law, regulation and technology*. Oxford Handbooks. Oxford University Press, London, England.

Daniel Carpenter and David A Moss, editors. 2013. *Preventing regulatory capture*. Cambridge University Press, Cambridge, England.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

European Commission. 2023a. Artificial intelligence – ethical and legal requirements. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/feedback_en?p_id=24212003. [Accessed 16-Jun-2023].

European Commission. 2023b. Better regulation: why and how. https://commission.europa.eu/law/law-making-process/planning-and-proposing-law/better-regulation_en#:~:text=The%20Better%20Regulation%20agenda%20ensures,those%20that%20may%20be%20affected. [Accessed 16-Jun-2023].

E Dal Bo. 2006. Regulatory capture: A review. *Oxf. Rev. Econ. Pol.*, 22(2):203–225.

Leon Derczynski, Hannah Rose Kirk, Vidhisha Balachandran, Sachin Kumar, Yulia Tsvetkov, MR Leiser, and Saif Mohammad. 2023. Assessing language model deployment with risk cards. *arXiv preprint arXiv:2303.18190*.

Roel Dobbe, Thomas Krendl Gilbert, and Yonatan Mintz. 2021. Hard choices in artificial intelligence. *Artif. Intell.*, 300(103555):103555.

Anjalie Field, Amanda Coston, Nupoor Gandhi, Alexandra Chouldechova, Emily Putnam-Hornstein, David Steier, and Yulia Tsvetkov. 2023. Examining risks of racial biases in NLP tools for child protective services. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, New York, NY, USA. ACM.

Luciano Floridi. 2023. *The ethics of artificial intelligence the ethics of artificial intelligence*. Oxford University Press, London, England.

Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds Mach.*, 30(3):411–437.

Megan Garber. 2012. The IVF Panic: 'All Hell Will Break Loose, Politically and Morally, All Over the World' — theatlantic.com. https://www.theatlantic.com/technology/archive/2012/06/the-ivf-panic-all-hell-will-break-loose-politically-and-morally-all-over-the-world/258954/. [Accessed 16-Jun-2023].

Audley Genus and Andy Stirling. 2018. Collingridge and the dilemma of control: Towards responsible and accountable innovation. *Res. Policy*, 47(1):61–69.

Catalina Goanta and Sofia Ranchordás. 2020. *The regulation of Social Media Influencers*. Edward Elgar Publishing.

Janet Gough. 1988. Risk and uncertainty. [Accessed 16-Jun-2023].

Jeroen Heijden. 2021. Why meta-research matters to regulation and governance scholarship: An illustrative evidence synthesis of responsive regulation research. *Regul. Gov.*, 15(S1).

Thomas A Hemphill. 2020. "the innovation governance dilemma: Alternatives to the precautionary principle". *Technol. Soc.*, 63(101381):101381.

Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. 2023. An overview of catastrophic ai risks.

Veerle Heyvaert. 2011. Governing climate change: Towards a new paradigm for risk regulation. *Mod. Law Rev.*, 74(6):817–844.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Lang. Linguist. Compass*, 15(8):e12432.

Anita Howarth. 2013. A 'superstorm': when moral panic and new risk discourses converge in the media. *Health Risk Soc.*, 15(8):681–698.

Liselot Hudders, Steffi De Jans, and Marijke De Veirman. 2021. The commercialization of social media stars: a literature review and conceptual framework on the strategic use of social media influencers. *Int. J. Advert.*, 40(3):327–375.

Celeste Kidd and Abeba Birhane. 2023. How AI can distort human beliefs. *Science*, 380(6651):1222–1223.

Jihyun Kim and Hayeon Song. 2016. Celebrity's self-disclosure on twitter and parasocial relationships: A mediating role of social presence. *Comput. Human Behav.*, 62:570–577.

John R. Krebs. 2011. Handling uncertainty in science. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, (1956):4842–4852.

Benjamin Laufer, Sameer Jain, A. Feder Cooper, Jon Kleinberg, and Hoda Heidari. 2022. Four years of facct: A reflexive, mixed-methods analysis of research contributions, shortcomings, and future prospects. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 401–426, New York, NY, USA. Association for Computing Machinery.

Johann Laux, Sandra Wachter, and Brent Mittelstadt. 2023. Trustworthy artificial intelligence and the european union AI act: On the conflation of trustworthiness and acceptability of risk. *Regul. Gov.*

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. 2018. Scalable agent alignment via reward modeling: a research direction. *CoRR*, abs/1811.07871.

Michael E. Levine and Jennifer L. Forrence. 1990. Regulatory capture, public interest, and the public agenda: Toward a synthesis. *Journal of Law, Economics, & Organization*, 6:167–198.

Gary E. Marchant, Braden R. Allenby, and Joseph R. Herkert. 2013. *The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight: The Pacing Problem*. Springer Publishing Company, Incorporated.

Fatemehsadat Mireshghallah, Kartik Goyal, Archit Uniyal, Taylor Berg-Kirkpatrick, and Reza Shokri. 2022. Quantifying privacy risks of masked language models using membership inference attacks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8332–8347, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Bronwen Morgan and Karen Yeung. 2007. *An introduction to law and regulation: Text and materials*. Cambridge University Press, Cambridge, England.

Anthony Ogus. 2009. Regulation revisited. *Public Law*, (2):332–346.

OpenAI. 2023. Gpt-4 technical report.

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

Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and David Moher. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Syst. Rev.*, 10(1):89.

K Francis Park and Zur Shapira. 2017. Risk and uncertainty. In *The Palgrave Encyclopedia of Strategic Management*, pages 1–7. Palgrave Macmillan UK, London.

Melanie Paschke, Andrea Pfisterer, Christian Hirschi, Luisa Last, Daniela Pauli, Bruno Studer, Jasmin Schubert, Robert Herrendörfer, and Kaitlin Elyse Mc Nally. 2019. Evidence-based policymaking.

Theodore M Porter. 2020. Objectivity and the politics of disciplines. In *Trust in Numbers*, pages 193–216. Princeton University Press.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Martin J Riedl, Josephine Lukito, and Samuel C Woolley. 2023. Political influencers on social media: An introduction. *Soc. Media Soc.*, 9(2):205630512311779.

Matthias C Rillig, Marlene Ågerstrand, Mohan Bi, Kenneth A Gould, and Uli Sauerland. 2023. Risks and benefits of large language models for the environment. *Environ. Sci. Technol.*, 57(9):3464–3466.

David Rolnick, Priya L. Donti, Lynn H. Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, Alexandra Sasha Luccioni, Tegan Maharaj, Evan D. Sherwin, S. Karthik Mukkavilli, Konrad P. Kording, Carla P. Gomes, Andrew Y. Ng, Demis Hassabis, John C. Platt, Felix Creutzig, Jennifer Chayes, and Yoshua Bengio. 2022. Tackling climate change with machine learning. *ACM Comput. Surv.*, 55(2).

Kevin Roose. 2023. A.I. Poses 'Risk of Extinction,' Industry Leaders Warn — nytimes.com. https://www.nytimes.com/2023/05/30/technology/ai-threat-warning.html. [Accessed 04-Jun-2023].

Andrea Saltelli, Dorothy J Dankel, Monica Di Fiore, Nina Holland, and Martin Pigeon. 2022. Science, the endless frontier of regulatory capture. *Futures*, 135(102860):102860.

Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. 2020. Green AI. *Commun. ACM*, 63(12):54–63.

Philip Selznick. 1985. Focusing organizational research on regulation. *Regulatory policy and the social sciences*, 1(1):363–367.

Felice Simonelli and Nadina Iacob. 2021. Can we better the european union better regulation agenda? *European Journal of Risk Regulation*, 12(4):849–860.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

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

Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla, and Oskar Van Der Wal. 2022. You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 26–41, virtual+Dublin. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Dimitrios Tsarapatsanis and Nikolaos Aletras. 2021. On the ethical limits of natural language processing on legal text. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3590–3599, Online. Association for Computational Linguistics.

David Valle-Cruz, J Ignacio Criado, Rodrigo Sandoval-Almazán, and Edgar A Ruvalcaba-Gomez. 2020. Assessing the public policy-cycle framework in the age of artificial intelligence: From agenda-setting to policy evaluation. *Gov. Inf. Q.*, 37(4):101509.

Jeroen van den Hoven. 2014. Responsible innovation: A new look at technology and ethics. In *Responsible Innovation 1*, pages 3–13. Springer Netherlands, Dordrecht.

Jose van Dijck. 2013. *The culture of connectivity*. Oxford University Press, New York, NY.

José van Dijck and Donya Alinejad. 2020. Social media and trust in scientific expertise: Debating the covid-19 pandemic in the netherlands. *Soc. Media Soc.*, 6(4):205630512098105.

Phillip Vannini and Alexis Franzese. 2008. The authenticity of self: Conceptualization, personal experience, and practice. *Sociol. Compass*, 2(5):1621–1637.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010, Long Beach, California, USA.

Michael Veale, Kira Matus, and Robert Gorwa. 2023. Ai and global governance: Modalities, rationales, tensions. *Annual Review of Law and Social Science*, 19.

Peter C Verhoef, Thijs Broekhuizen, Yakov Bart, Abhi Bhattacharya, John Qi Dong, Nicolai Fabian, and Michael Haenlein. 2021. Digital transformation: A multidisciplinary reflection and research agenda. *J. Bus. Res.*, 122:889–901.

René von Schomberg. 2013. A vision of responsible research and innovation. In *Responsible Innovation*, pages 51–74. John Wiley & Sons, Ltd, Chichester, UK.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 214–229, New York, NY, USA. Association for Computing Machinery.

Robert Wolfe, Yiwei Yang, Bill Howe, and Aylin Caliskan. 2023. Contrastive language-vision ai models pretrained on web-scraped multimodal data exhibit sexual objectification bias. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1174–1185, New York, NY, USA. Association for Computing Machinery.

Jennifer Yarnold, Ray Maher, Karen Hussey, and Stephen Dovers. 2022. Uncertainty. In *Routledge Handbook of Global Environmental Politics*, pages 253–268. Routledge, London.

Annie Li Zhang and Hang Lu. 2023. Scientists as influencers: The role of source identity, self-disclosure, and anti-intellectualism in science communication on social media. *Social Media + Society*, 9(2):20563051231180623.

Yini Zhang, Chris Wells, Song Wang, and Karl Rohe. 2018. Attention and amplification in the hybrid media system: The composition and activity of donald trump's twitter following during the 2016 presidential election. *New Media Soc.*, 20(9):3161–3182.