Journal of Intellectual Property Studies
Vol. 4 (01), January 2021, pp. 71-91

THE LEGAL DIALECT OF NLP: *BHARAT NLP& AI FOR ALL*

SHARAN A. BHAVNANI*

ABSTRACT

Information & Communications Technologies (ICT) have displayed the potential to act as an agent for social and economic development. There are two aspects to enabling such development. The first, access to ICT; second, access through ICT. Thus, to empower the poor and help alleviate poverty, ICT must not only be easily available but also make accessible other existing information and knowledge. Recognising this, NITI Aayog released the National Strategy Paper on Artificial Intelligence which lays out the Bharat Natural Language Processing Program. This program accounts for the vibrant and pluralistic nature of India and establishes a rudimentary framework of access to and through ICT by the use of various vernacular languages. This framework is assessed by first, exploring the Indian legal framework, and second, the international framework to unpack India's international obligations. The domestic legal architecture of intellectual property rights has gaps in accounting for software that deploys Machine Learning and Natural Language Processing requiring vast corpora of datasets. The paper makes the case for viewing the use of such datasets within the "fair dealing" exception enumerated in the Indian Copyright Act, 1957. It also highlights the limitations of the Indian legal framework and makes policy recommendations for the commercial use these vernacular corpora of data for the Bharat NLP Program, to overcome these limitations in the short and long term.

*Sharan A Bhavnani, National Law School of India University, Bangalore. The author can be reached at sharanb@nls.ac.in

I. INTRODUCTION

More than mere material deprivation, poverty is a product of one's exclusion from the access to information and knowledge.[1] With the exponential growth of Information & Communication Technologies ["**ICTs**"], an undeniable nexus with alleviating poverty has emerged.[2]It has the potential to foster "broader developmental impact, empowerment and income generation, and increase access to education and other social services."[3] In recognition of the same, the Hon'ble Prime Minister of IndiaShri. Narendra Modi has expressed that "technology empowers the less empowered. If there is a strong force that brings a change in the lives of those on the margins , it is technology."[4]There are then two aspects of access to information and knowledge. The first, access *to* ICT; Second, the access *through* ICT. Thus, to empower the poor and help alleviate poverty, ICT must not only be easily available but also make accessible other existing information and knowledge.

The National Strategy Paper on Artificial Intelligence by NITI Aayog displays data which finds that:

> "*Indian language internet users are projected to account for nearly 75% of India's internet user base by the year 2021. However, the amount of digital content available in Indic regional languages is meagre compared to the expected demand: only 0.1% of internet content is in Indic languages vastly in Hindi, and the figure is far lower for dynamic contents and non-recreational resources such as news or education-oriented applications.*"[5]

To further the agenda of a Digital India and use of technology to empower those who "live on the margins" of society, the NITI Aayog ["**NITI**"] is tasked with designing an Artificial Intelligence ["**AI**"] framework, which is accessible to all Indians, irrespective of the language they speak. It seeks to achieve a part of this through the Bharat Natural Language Processing

---

[1]Caroline M. Figuères, & Eugelink, Hilde, *The Role of ICTs in Poverty Eradication: More Than 15 Years' Experience from the Field*, http://doi-org-443.webvpn.fjmu.edu.cn/10.1007/978-1-4899-7439-6_12.

[2]J. May, *The ICT/Poverty Nexus*, THE UN CHRONICLE (Vol. XLVIII) (3) (2011) https://www.un.org/en/chronicle/article/ictpoverty-nexus.

[3]*Id.*

[4] Press Release, Narendra Modi, Prime Minister, *Digital Dialogue with PM Narendra Modi*, PM-INDIA WEBSITE July 5, 2015,http://www.pmindia.gov.in/en/news_updates/digital-dialogue-with-pm-narendra-modi/?tag_term=digital-india&comment=disable.

[5] NITI Aayog, *National Strategy for Artificial Intelligence*, INTERNATIONAL INNOVATION CORPS AND UNIVERSITY OF CHICAGO (2018),http://niti.gov.in/writereaddata/files/document_publication/NationalStrategy-for-AI-Discussion-Paper.pdf.

["**NLP**"] and Machine Learning ["**ML**"].[6] The Bharat NLP intends to acquire and process numerous datasets in Indic languages to bridge the increasing asymmetry between usage of Indic languages, and the availability of the same in software and technology. To form this framework, NITI intends to make indigenous and regional languages accessible to tech-based start-ups, computer science and linguistics researchers, and larger technology companies through NLP tools. In its pursuit of the same, NITI would require vast corpora of text and information to create effective NLP tools.

There are three "building blocks"[7] of the Bharat NLP, namely, *first*, the web portal or repository that will work as a one-stop-shop for NLP resources for Indic languages; *second* a set of NLP tools to enable researchers and developers to provide digital resources in regional languages, and *third*, the language corpora datasets necessary to develop such NLP tools.

This article explores the social impact and legality of procuring and using these vast corpora of data sets by looking at successful pilot programmes, the relationship between rights and accessibility of rights, the legal position in India concerning Intellectual Property Rights, and policy recommendations to achieve the goal of a Digital India which is accessible to all.

II. ICT, ACCESSIBILITY AND SOCIAL GROWTH

The Indian Constitution expresses the aspirations of the people through the language of directive principles. The *first section* of this part explores these in the context of accessibility to and through ICT. The *second section* seeks to explore how these aspirations translate into governmental and social action. The *third section* delves into similar actions in other parts of the globe and emergent successes of using ICT for social growth.

A. CONSTITUTIONAL ASPIRATIONS AND RECOGNITIONS

The ever-growing borderless technologies have the potential to cause vast social inequalities and entrench inequities. Highlighting emerging challenges to the Right to Privacy in a globalised and technologically woven world, Hon'ble Dr. Justice D.Y. Chandrachud opined that,

---

[6] Li Deng, *et al.*, DEEP LEARNING IN NATURAL LANGUAGE PROCESSING (Deng, Li, Liu, Yang eds., 2018),("Natural language processing (NLP) investigates the use of computers to process or to understand human (i.e., natural) languages for the purpose of performing useful tasks. NLP is an interdisciplinary field that combines computational linguistics, computing science, cognitive science, and artificial intelligence…NLP aims to model cognitive mechanisms underlying the understanding and production of human languages."); See also, Ian Goodfellow *et al.*, DEEP LEARNING, 2 (2016),("The difficulties faced by systems relying on hard-coded knowledge suggest that AI systems need the ability to acquire their own knowledge, by extracting patterns from raw data. This capability is known as machine learning.")

[7] NITI Aayog, *supra* note 5.

> "*The overarching presence of state and non-state entities regulates aspects of social existence which bear upon the freedom of the individual. The preservation of constitutional liberty is, so to speak, work in progress. Challenges have to be addressed to existing problems. Equally, new challenges have to be dealt with in terms of a constitutional understanding of where liberty places an individual in the context of social order. The emergence of new challenges is exemplified by this case, where the debate on privacy is being analysed in the context of a global information-based society. In an age where information technology governs virtually every aspect of our lives, the task before the Court is to impart constitutional meaning to individual liberty in an interconnected world.*"[8]

However, the use of ICT as a tool, by the State for social development rather than letting it be a medium for the same, causes inequality and this is guided by the Directive Principles of State Policy, where the State,

> "shall strive to promote the welfare of the people by securing and protecting as effectively as it may a social order in which justice, social, economic and political, shall inform all the institutions of the national life. In particular, *strive to minimise the inequalities in income, and endeavour to eliminate inequalities in status, facilities and opportunities, not only amongst individuals but amongst groups of people residing in different areas or engaged in different vocations*."[9] (emphasis added)

Harold Laski in his notable work, 'The Grammar of Politics', states that "the freedoms I must possess to enjoy a general liberty are those which, in their sum, will constitute the path through which my best self is capable of attainment."[10] The path to the "best self" may only be attained when the path is paved with general liberties and rights; but more importantly, this path must be accessible to all to exercise these general liberties. This is only possible when one has the accessibility to mechanisms that enhance rights and social growth.

B. TRANSLATING CONSTITUTIONAL ASPIRATIONS INTO POLICIES AND ACTION

[8] Justice K.S. Puttaswamy (Retd.) v. Union of India, (2014) 6 SCC 433, Opinion of Justice D.Y. Chandrachud.
[9] Constitution of India, art. 38 (1950).
[10] HAROLD J. LASKI, A GRAMMAR OF POLITICS, 144 (1925).

In 1995, Heads of States and Government from 117 nations emphasised the role of new information technologies in poverty alleviation and realisation of social developmental goals in Copenhagen; recognising the need for countries to facilitate access to such technologies.[11]

With the leaps made in information technology since 1995, the Government of India in 2014 conceived the broad goal of creating a Digital India. The Government – guided by Article 38 of the Indian Constitution and its recognition of the role of ICT in social growth – has initiated the linking of various subsidies through the implementation of the Aadhar programme at the very grassroots.[12]

This has made the system of the Government efficient and ensures the timely and effective use of provisions of social benefits through the use of ICTs. Within this broader agenda of a Digital India is the Government's concentration on emergent technologies such as AI, which can make ground-breaking changes at the grassroots level in ensuring social growth in several aspects.[13]

Uncontestably, the use of ICT in a broad gamut of languages, which are accessible to all, can lead to the vast social and educational development of marginalised communities.[14] Where "recent developments in Information and Communications Technologies (ICTs) can make to the management of multilingualism and how local languages can be used to make education a lever of development"[15] is only one of the ways to use ICTs to fillip development, including educational development.

That being said, most information available and/or programmes developed through ICT for education and socio-economic growth is available only in English.[16] India being linguistically diverse must necessarily make access to existing and new technologies through a multilingual

---

[11] Rep. of the World Summit for Soc. Dev., UN DOC A/CONF.166/9 (1995), ("States in their Copenhagen Declaration on Social Development and Programme of Action of the World Summit for Social Development went on to, "Recognize that the new information technologies and new approaches to access to and use of technologies by people living in poverty can help in fulfilling social development goals; and therefore recognize the need to facilitate access to such technologies").

[12] Ministry Of Consumer Affairs, Food & Public Distribution, *Linking of Aadhaar with PDS*, PRESS INFORMATION BUREAU, Feb. 4, 2020, https://pib.gov.in/PressReleasePage.aspx?PRID=1601866.

[13] Michael Chui *et al.*, *Applying artificial intelligence for social purposes: Discussion Paper*, MCKINSEY GLOB. INST.,(2018), https://www.mckinsey.com/featured-insights/artificial-intelligence/applying-artificial-intelligence-for-social-good.

[14] Paulin. G. Djité, *The nexus between education, learning, and language*, UNESCO CONFERENCE ON "GLOBALIZATION AND LANGUAGES: BUILDING ON OUR RICH HERITAGE" (2008), http://archive.unu.edu-/globalization/2008/files/UNU-UNESCO_Dijite.pdf.

[15] *Id.*

[16] KPMG India & Google, *Indian Languages- Defining India's Internet*,(2017) https://assets.kpmg/content/dam-/kpmg/in/pdf/2017/04/Indian-languages-Defining-Indias-Internet.pdf.

platform. This has been recognised by the Government with the Hon'ble Minister of Electronics and Information Technology Shri. Ravi Shankar Prasad stating "the language of the internet cannot be English and English alone. It must have linkages with the local and local means local languages. I appeal to make local languages available for more Internet users."[17]

The Government has recognised the need for ICT tools for seamless translation from English to vernacular languages. This can be seen in the establishment of the MANTRA software by the Department of Language, which allows the accurate translation of English documents into Hindi using NLP. This NLP programme is used by "various departments and ministries of the government to procure quick and standard Hindi translations of English documents." [18] The contextual nature of the translation has made these documents available to a large segment of Hindi speakers across the nation, let alone the institutions of the Government thereby increasing efficiency in Government functioning.

Another such instance in India is the use of NLP and Optical Character Recognition to provide the visually challenged with audio descriptions of their immediate surroundings.[19] The programme, DRISHTI, pioneered by Accenture and the National Association of Blind in India, makes accessibility to the visually challenged easier, which is a right guaranteed under the Constitution of India and various legislations.[20] This programme is also being developed in Spanish for its use in Argentina.[21] The use of such technology in India has also been deployed by inter-governmental agencies like the World Bank.

For inclusive growth through deliberation and participation by women in forums of local decision-making, the World Bank conducted a study in 2017, in Indian villages using NLP to assess the quality of participative democracy. The purpose of the study was to explore "whether and to what extent…forms of dominance — of men over women… affect deliberative institutions in practice" In order to assess this on a large scale, the study used "Natural Language Processing (NLP, or text-as-data) methods to an original corpus of village assembly transcripts

---

[17] P. Hegde, *Protecting Language Diversity in India*, PRESS INFORMATION BUREAU, Govt. of India, (2017), http://pib.nic.in/newsite/mbErel.aspx?relid=158532.

[18] Centre for Development and Advanced Computing MANTRA Rajbhasha, available at: https://www.cdac.in/index.aspx?id=mc_mat_mantra_rajbhasha.

[19] *Defying Limits*, ACCENTURE LABS ANNUAL REPORT (2017) https://www.accenture.com/t20180221T1937-48Z__w__/us-en/_acnmedia/PDF-70/Accenture-Labs-2017Report-Digital.pdf.

[20] Convention on the Rights of Persons with Disabilities, Dec. 13, 2006 , UNITED NATIONS A/RES/61/106 (2006).

[21] Accenture News Release, *Accenture Develops Artificial Intelligence-Powered Solution to Help Improve How Visually Impaired People Live and Work* July 28, 2017, https://newsroom.accenture.com/news/accenture-develops-artificial-intelligence-powered-solution-to-help-improve-how-visually-impaired-people-live-and-work.htm.

from rural India to systematically examine variation in the quality of deliberation [and] examine the relationship between deliberative influence and the gender or position (citizen versus official) of a speaker."Using NLP methods, the study could "quantitatively examine not only the relative floor time enjoyed by different types of speakers but also their ability to influence the topic of conversation (agenda-selling power) and to make claims on state officials (responsiveness of the state)." With these ICT methods emerged a clear depiction of the challenges faced in the grassroots governance. The study found that despite the high rates of attendance of women in village-assemblies, they were "*indeed the silent sex."* This was seen through the disparity between their average attendance (58%) and the available floor time (33%). It also found that "when women do speak on a particular topic, they are significantly less likely than men to elicit a topical or relevant response from state officials — suggesting a meaningful inequality in deliberative influence across the sexes....for any given topic, a man is more likely to get a response from an official than a woman."[22] These findings provide valuable insight into the efficacy of democratic institutions and the extent of women's participation in a participatory democracy.[23]

C. ENCOURAGING GLOBAL DEVELOPMENTS

The use of ML and NLP has shown encouraging results, not only in Indian Governmental institutions (to a limited extent) but also across different parts of the globe.

For instance, students from the Jomo Kenyatta University of Agriculture and Technology (JKUAT) in Kenya developed the "SophieBot".[24] This application bridges the information asymmetry by translating information on women's sexual and reproductive health, from English to Swahili. This also helped in further disseminating vital information about health and reproductive rights in Kenya, thereby promoting a favourable environment for the protection of human rights and the realisation of Sustainable Developmental Goals.[25]

---

[22] Ramya Parthasarthy*et al.*, *Deliberative Inequality: A Text-As-Data Study of Tamil Nadu's Village Assemblies*, 2-3, WORLD BANK GROUPPOL'Y RES. WORKING PAPER NO. 8119, 2017), http://documents.worldbank.org/curated-/en/582551498568606865/pdf/WPS8119.pdf.

[23]The Convention on the Elimination of all Forms of Discrimination Against Women(CEDAW), UNITED NATIONS A/RES/34/180 (1981).

[24] Department of Information Technology, *Sophie Bot Sexual Reproductive Health app*, JOMO KENYATTA UNI VE RSI TY OF AGRICULTURE AND TECHNOLOGY,http://www.jkuat.ac.ke/departments/it/sophie-bot-sexual-rep roductive-health-app/.

[25]*UNFPA Announces Winners of Innovation Accelerator Focused On Promoting Youth Sexual Reproductive Health and Rights*, UNITED NATIONS POPULATION FUND AGENCY (UNFPA), AUG. 12, 2016, https://kenya.unfpa.org/en/news/unfpa-announces-winners-innovation-accelerator-focused-promoting-youth-sexual-reproductive.

In Nigeria, a start-up using ML and NLP developed software called "Kudi.AI" which makes e-banking and digital payments accessible to many people in English and to an extent in Pidgin (the native language) through the use of NLP to contextualise the queries and actions by users.[26] This comes in the wake of Governmental support along with the Presidential approval for the creation of an AI and Robotics Authority.[27] This entrepreneurial drive along with the broad framework provided by the Government makes for inclusive and equitable growth of a country.

These are only a few examples of the use of ML and NLP, which have enhanced access to various socio-economic and civil rights. Various other avenues can be traversed with the use of AI to achieve numerous Sustainable Developmental Goals.[28]

In the implementation of AI such as ML and NLP in India, especially the acquisition of vast swathes of data and text could lead to certain domestic legal challenges. This shall be explored in the next section through analysing the intersection of rights and accessibility of rights – within the India and International frameworks. Further, this section shall analyse the specific copyright issues within India and ultimately make policy recommendations.

III. RIGHTS AND ACCESSIBILITY

People have the right to hold aspirations and to acquire information and knowledge. Given the culturally vibrant and pluralistic nature of India, it is important to assess the efficacy of access to and through ICT via the use of various vernacular languages. This shall be assessed by *first*, exploring the Indian legal framework which makes clear these aspirations, and *second*, the international framework to unpack the kinds of obligations a State must follow.

A. INDIAN FRAMEWORK

The Preamble to the Indian Constitution secures unto the citizens of India- Justice, Liberty, Equality and Fraternity to ultimately ensure social, economic, political, personal, and collective growth.[29] The conception of justice, equality of opportunity and fraternity ensuring the unity and integrity of the nation can only be attained when there is unfettered access to the rights and ideals guaranteed under our Constitution. One aspect is accessibility to institutions of the State to

[26] KUDI WEBSITE, available at: https://kudi.co/.

[27] Nnamdi Akpa and Abakaliki, *Buhari approves artificial intelligence agency for South East*, THE GUARDIAN Aug. 6, 2018, https://guardian.ng/news/buhari-approves-artificial-intelligence-agency-for-south-east/.

[28] Chui, *supra* note 13.

[29] Constitution of India, Preamble (1950), ("The Constitution guarantees unto citizens, Justice, Social, Economic & Political; Liberty of thoughts, expression, belief, faith and worship; Equality of status and of opportunity and promote among them all; Fraternity, assuring the dignity of individual and the unity and integrity of the nation.").

secure justice and enforce rights for a dignified life.[30] Another aspect is the right to information and freedoms of speech and expression, which is accessible to all. In addition to the nexus between ICT and social development, there lies a nexus between the accessibility of rights and language.[31]

The Supreme Court of India, in interpreting Article 19(1)(a) observed that the provision aims at "promoting the dissemination of ideas, information and knowledge to the masses so that there may be an informed debate and decision making on public issues."[32] Naturally, this dissemination is only effective when it is available to all, irrespective of the language they speak in. A step towards the creation of this accessibility through various means and languages can be seen in the multitude of legislative and policy measures undertaken by the state. For instance, the Right to Information Act, 2005 specifies in Section 4 that "All materials shall be disseminated taking into consideration the local language"

Another instance is the development of the National Policy on Universal Electronic Accessibility by the Ministry of Electronics and Information and Technology, in 2007.[33] The policy is based on barriers which differently-abled individuals face. This policy, grounded in the Constitution of India and guided by several International covenants,[34] created enough traction to lead to an Amendment in the Copyright Act, 1957 in the year 2012.[35] The amendment created an exception for braille or any other format, which would help the visually impaired in accessing any work with copyright.[36] We see that the National Language Policy seeks to be "pluralistic in its scope" and intends to "help all languages to develop into fit vehicles of communication at their

---

[30] Anita Kushwaha and Ors. v. Pushap Sudan, AIR 2016 SC 3506 (India).

[31] T. Skutnabb-Kangas, *Role of Linguistic Human Rights in Language Policy and Planning*, THE ENCYCLOPAEDIA OF APPLIED LINGUISTICS (C. Chapelle ed.,2012), DOI: 10.1002/9781405198431.wbeal1026.

[32] Union of India v. Motion Pictures Association, (1999) AIR 1999 SC 2334 (India).

[33] Ministry of Electronics and Information Technology, *National Policy on Universal Electronic Accessibility*, (2007), https://meity.gov.in/writereaddata/files/National%20Policy%20on%20Universal%20Electronics%281%29.pdf.

[34] *Id.* at 2 ("The United Nations Convention for the Rights of Persons with Disabilities ratified by India on October 01, 2007 (UNCRP D - Article 9 and others). (iii) Persons with Disabilities (Equal Opportunities, Protection of Rights and Full Participation) Act, 1995. (iv) The Proclamation on the Full Participation and Equality of People with Disabilities in the Asia Pacific Region, 1993. (v) The Biwako Millennium Framework for action towards an inclusive, barrier free and rights based society, 2002. (vi) National Policy for Persons with Disabilities of the Government of India (2006).").

[35] Copyright (Amendment) Act, 2012, No. 27, Acts of Parliament, 2012 (lndia).

[36] Copyright Act, No. 14 of 1957, §52(zb), (India).

designated areas of use, irrespective of their nature or status like major, minor, or tribal languages."[37]

The role of languages as a facilitator rather than a barrier helps in achieving Constitutional guarantees and socio-economic growth, and is also visible within the Constitutional framework. All persons are entitled to make a representation to any authority or officer of the Union or State in "any languageused in the Union or State"[38] to have their grievances redressed. Further, regarding linguistic minorities, all States and local authorities shall endeavour to "provide adequate facilities for instruction in the mother-tongue at the primary stage of education to children belonging to linguistic minority groups"[39] and the President appoints a Special Officer for Linguistic Minorities who is tasked with the safeguarding of such minorities.[40]

With 22 Official Languages, and one Associate Official Language in Schedule VIII of the Indian Constitution, the plurality and linguistic diversity make up a truly federal India, the legal and policy framework makes it clear that the language is essential for access to rights and enforcement of the same. India, like many nations with linguistic diversity, is a party to various international covenants and declarations which reinforces this link.

B. INTERNATIONAL FRAMEWORK

This sub-section shall delve into, *first*, India's obligations within the International legal framework; and *second*, the case for protection and promotion of the rights guaranteed under various treaties or customary norms.

India is a party to various treaties, which guarantees non-discrimination based on language. Further, there are various other treaties which guarantees a broad gamut of human rights. These rights may only be enjoyed if and when the rights are accessible, notwithstanding the barriers of language. For instance, as a part of Customary International law,[41] the Universal Declaration on Human Rights ["**UDHR**"] expresses shared values of states in the protection and promotion of Human Rights. These rights should not be limited by discrimination on the basis of "language".[42]Another instance is that India is a signatory to the Declaration on the Right to

---

[37] Ministry of Human Resource Development, *Overview: Language Policy of India*, GOVERNMENT OF INDIA http://mhrd.gov.in/language-education.

[38]Constitution of India, art. 350 (1950).

[39]*Id.* at art. 350A.

[40]*Id.* at art. 350B.

[41]A. D'AMATO, INTERNATIONAL LAW: PROCESS AND PROSPECT123-147 (1987).

[42]*Id.* at art. 2.

Development, 1986 DRD), and cannot derogate from its core principles,[43] especially "the primary responsibility for the creation of national…conditions favourable to the realisation of the right to development."[44]

Given that these rights and principles are imperative for the delivery of essential human rights; they are to be read broadly in so far as the obligation they impose on the state. The International Law Commission, which drafted the Responsibility of States for Internationally Wrongful Acts ["**ARISWA**"][45] lays down the meaning of "breach"[46] and the nature of the obligation.[47] A breach takes place when a state is not in conformity with what is required by it. In his Commentary of the ARISWA, (now Judge) James Crawford notes that the "phrase "not in conformity with" is flexible enough to cover the many different ways in which an obligation can be expressed, as well as the various forms which a breach may take."[48]To meet the intended effectiveness of human rights obligations, when construed broadly, their nature is that of result, unless the obligation explicitly provides otherwise. The language of DRD, several human treaties and customary law make clear that it is not only the mere obligation to create a law or mechanism to protect rights, but also to promote an environment which ensures these rights are effective and accessible.

It is through the State's intervention that a bulk of human rights obligations can be achieved. When viewed from the lens of the obligation of result, i.e. obligations which truly seek to

---

[43] DECLARATION ON THE RIGHT TO DEVELOPMENT (DRD), UNITED NATIONS A/RES/41/128 (1986). [hereinafter "Declaration, 1986"], https://www.un.org/en/events/righttodevelopment/declaration.shtml.
[Art. 1 (1) states "*The right to development is an inalienable human right by virtue of which every human person and all peoples are entitled to participate in, contribute to, and enjoy economic, social, cultural and political development, in which all human rights and fundamental freedoms can be fully realized…*"
(Art. 3 (1) states, "*States have the primary responsibility for the creation of national and international conditions favourable to the realization of the right to development.*"
Art. 6 states, "*1. All States should co-operate with a view to promoting, encouraging and strengthening universal respect for and observance of all human rights and fundamental freedoms for all without any distinction as to race, sex, language or religion. 2. All human rights and fundamental freedoms are indivisible and interdependent; equal attention and urgent consideration should be given to the implementation, promotion and protection of civil, political, economic, social and cultural rights. 3. States should take steps to eliminate obstacles to development resulting from failure to observe civil and political rights, as well as economic social and cultural rights.*"
Art. 8 (1) states, "*States should undertake, at the national level, all necessary measures for the realization of the right to development and shall ensure, inter alia, equality of opportunity for all in their access to basic resources, education, health services, food, housing, employment and the fair distribution of income*.").

[44] Declaration, 1986, *Id*, art. 1.

[45] RESPONSIBILITY OF STATES FOR INTERNATIONALLY WRONGFUL ACTS (ARSIWA), UNITED NATIONS A/RES/56/83 (2001). [hereinafter "**ARSIWA, 2001**"]

[46] *Supra* note 45 art. 12.

[47] *Supra* note 45 art. 21.

[48] *Commentary to the ARSIWA*, U.N. DOC. A/CN.4/SER.A/2001/Add.1 (Part 2) [2001] Y.B. INT'L L. COMM'N Vol. II (2) 55), https://legal.un.org/ilc/publications/yearbooks/english/ilc_2001_v2_p2.pdf.

empower the marginalised (socially, economically, politically, and culturally), the state must create a conducive environment to achieve this. Its non-conduct of the creation of such an environment would lead to a breach of these norms.[49] Undeniably, there is a link between human rights, their accessibility, and the use of ICT in ensuring this conducive environment so that India may meet its obligations under International law.

IV. DATASETS & COPYRIGHT

The creation of software that deploys ML and NLP requires vast corpora of datasets. Unfortunately, the availability of the same in Indic languages may be difficult for two reasons. *First*, there is a paucity of text available in Indic languages online, and *second*, that the available data is protected by copyright. This section shows how the use of such datasets falls within the "fair dealing" exception enumerated in the Indian Copyright Act, 1957. This section shall also highlight the limitations of the Indian legal framework, and make policy recommendations for the use of such text for the Bharat NLP programme.

A. DATASETS AVAILABLE

To ensure that the NLP and ML programmes function to begin with and attain their optimum level of contextual accuracy, they need to be fed with vast corpora of data in Indic languages.

At present, there are various stakeholders within and outside the Government that possess textual data in various Indic languages.[50]This textual data helps AI and algorithms with

[49]Colozza v. Italy, (1985) Eur. Court H.R., 89 (ser. A), art28 (European Court of Human Rights); United Nations, *Chapter III: Breach of International Obligation*, MATERIALS ON THE RESPONSIBILITY OF STATES FOR INTERNATIONALLY WRONGFUL ACTS, UNITED NATIONS 102 (2012)("…it (the ECtHR) did notsimply compare the result required (the opportunity for a trial in the accused's presence) with the result practically achieved (the lack of that opportunity in the particular case). Rather, it examined what more Italy could have done to make the applicant's right "effective"…Here the Court looked at steps towards effectiveness of the implementation of a right as the standard."), https://legal.un.org/legislativeseries/pdfs/chapters/book25/book25_part1_ch3.pdf.

[50] The following is only an indicative list of these stakeholders: Department of Official Language, Ministry of Home Affairs (with various Hindi eTools, which include the NLP programme MANTRA and eMahashabdakhosh, the copyright for which is held jointly by the Department of Language and Centre for Development of Advanced Computing); Central Board of Secondary Education (CBSE) (with various literature books in Sanskrit, Urdu, Hindi, etc., the copyright for which is held by the CBSE Research & Development Unit and NCERT); Andhra Pradesh Board of Intermediate Education and the Andhra Pradesh Open School Society (with literature books in Telugu, the copyright for which is held by the Andhra Pradesh State Council for Education and Training); Assam Higher Secondary Education Council (with literature books in Assamese, Bodo and Bengali, the copyright of which is held by the Assam State Council for Education and Training); Council for Indian School Certificate Examinations (ISC) (with various texts in Hindi, Arabic, Sanskrit, Persian, Lepcha, Urdu, Telugu, Tamil, Punjabi, Odia, Marathi, Manipuri, Nepali, Khasi, Malayalam, Mizo, Kannada, Dzongkha, Gujrati. Bengali and Assamese, the copyright of which is held by ISC); The Federation of Indian Publishers (FIP) (They represent approximately 80% of the Indian publishers across India which publish in English, Hindi, and regional Languages. The copyright is held by

understanding and learning the rules of language and establishing context. The more the data, the better trained the algorithms would be.Acquisition of this data would be key in training the algorithms, which when made available to start-ups can be deployed for various purposes in industries like health, education, social empowerment, etc. and eventually increase access to basic Constitutional guarantees and human rights. There are broadly two types of datasets which may be acquired, those in and out of the public domain.

C. LEGAL OPTIONS AVAILABLE TO NITI AAYOG

This sub-section shall engage with the legality of obtaining and using these datasets using provisions of the Indian Copyright Act, 1957 and judicial interpretation of the same. Two primary arguments are made, *first*, there is no copyright infringement in instances where the data is in the public domain, and *second*, the use of data-sets not in the public domain is covered by the exception of "fair dealing". The third sub-section will address the limitations to the law as it stands.

1. NO INFRINGEMENT WHEN DATA IS IN THE PUBLIC DOMAIN

When a work is said to enter the public domain, it means that no copyright over it exists and can be used by anyone. Three key provisions of the Act lay down the period for the subsistence of a copyright over a literary work.[51]For literary works, the copyright subsists for 60 years after the death of the author (in the case of joint work, it is 60 years from the death of the author who dies last), for pseudonymous works it is 60 years from publication (unless the identity of the author comes to light, then it is the same number of years from the coming to light of the identity), and for posthumous works is it 60 years from the date of publication.

2. DATA NOT IN THE PUBLIC DOMAIN IS COVERED BY THE EXCEPTION OF FAIR DEALING

The copyright over various literary works is protected by section 14, Indian Copyright Act, 1957 as,

---

individuals, but hold deliberations within its own 'Copyright Council', which could be engaged for Bharat NLP); Data held by IITs (by individual IITs and some held with consortiums. Some of the data includes Neural POS Tagger (Neural NER) (IIT Bombay), Hindi English Parallel Corpus (IIT Bombay), Indic TTS (IIT Madras), Datasets for FIRE 2013 Track on Transliterated Search – consists of corpora with Hindi, English, Bengali, Gujarati (IIT Kharagpur), Bengali ASR Speech Corpus (IIT Kharagpur), Speech Database of Language Identification (IIT Kharagpur), Telugu Emotional Speech Corpus (IITKharagpur), Hindi Emotional Speech Corpus (IIT Kharagpur), and Speech Database for unsupervised clustering (IIT Kharagpur).

[51] Copyright Act, *supra* note 36, § 22, 23 & 24.

"copyright" means the exclusive right subject to the provisions of this Act, to do or authorise the doing of any of the following acts in respect of a work or any substantial part thereof, namely:

(a) in the case of a literary…work

(i) to reproduce the work in any material form <u>including the storing of it in any medium by electronic means</u>;"[52]

When NITI stores this data within its corpora for ML and NLP, the aforementioned section covers it.

The Act, however, carves out exceptions and broadly classifies the same as "fair dealing".[53] Reasonable restrictions within the Constitutional framework which manifests as exceptions within Section 52 of the Act are "keeping with the utilitarian public benefit theory of copyright law so that public access to content and its necessary dissemination is not curtailed by the rights granted to the author... thus, fair dealing is a right granted to the public under the Copyright Act"[54] The relevant part of the section reads as follows:

"Section 52. Certain acts not to be infringement of copyright.—

(1) The following acts shall not constitute an infringement of copyright, namely,—

(a) a fair dealing with any work, not being a computer programme, for the purposes of—

(i) private or personal use, including <u>research</u>;"

The High Court of Madras has interpreted the word "research" to mean, "an investigation directed to the discovery of some fact by careful study of a subject; investigation, inquiry into things."[55] The NITI with its Bharat NLP seeks to conduct a thorough investigation and inquiry to discover the nature and rulesof various Indic languages by having the software analyse the data corpora it acquires. Further, technology has grown exponentially since the inception of the Act and the Supreme Court of India has propounded a broad reading of the language in the law by opining that, "interpretation of every statutory provision must keep pace with the changing

[52] *Id.* §14(a)(i).
[53] *Id.* §52.
[54] R. Matthan *et al.*, *Fair Dealing of Computer Programs in India*, 7 INDIAN J. L. TECH. 93-94 (2011).
[55] Blackwood and Sons Ltd. v. A.N. Parasuraman, AIR 1959 Mad HC 410 (India).

concepts and it must, to the extent to which its language permits, or rather does not prohibit, suffer adjustments so as to accord with the requirements of fast-growing society."[56]

Natural Language Processing is an outcome of this "fast-growing society", and tools such as these, by learning the rules of various Indic languages, ultimately encourage the dissemination of various kinds of information through the creation of the Bharat NLP software.[57] Recognising the role of technology in the betterment of society, the Supreme Court of India has noted that "no law can be interpreted so as to result in any regression of the evolvement of the human being for the better."[58]

With such a focus from the Government, Indian society only moves forward as the grow ing need for Indic languages (many of which are otherwise on the verge of extinction) in upcoming technologies would be met through such focus. This would further enable accessibility to a mass of people who are excluded from availing the benefits of technological applications, which help in the awareness of basic rights. A restrictive reading would amount to "lowering the aspirations of public at the behest of publishers (also in the name of authors) by legislature or courts and would be the greatest disservice to the nation and the constitutional guarantees."[59] Therefore, the utilisation of data corpora of Indic literary works would come within the exception of "fair dealing" in the Act.

A counter-argument to the "fair dealing" exception may be based on the reliance on Section 51 of the Act which *deems* infringement in certain cases. The relevant part of the section reads as follows,

> "*51. When copyright infringed.— Copyright in a work shall be deemed to be infringed—*
>
> *(b) when any person—*
>
> *(ii) distributes either for the purpose of trade or to such an extent as to affect prejudicially the owner of the copyright,*

---

[56] S.P. Gupta v Union of India, (1982) AIR 1982 SC 149, Opinion of J. Bhagwati on p. 62 (India); The Chancellor, Masters and Scholars of the University of Oxford v. Rameshwari Photocopy Services p. 11 (2016) 233 DLT 279 (India).

[57] Matthan, *supra* note 58 at 94 reads "*...the various exemptions and doctrines implicit in copyright law, whether statutorily embedded or judicially innovated, recognize the equally compelling need to promote creative activity and ensure that the priv ileges granted by copyright do not stifle dissemination of information...*"; The Chancellor Masters and Scholars of the University of Oxford v. Narendera Publishing House, (2008) 106 D.R.J. 482, p. 32 (India).

[58] The Chancellor, Masters and Scholars of the University of Oxford v. Rameshwari Photocopy Services p. 87 (2016) 233 DLT 279 (India).

[59] A. K. Bansal, *Public Interest in Intellectual Property Laws*, 55(4) J. INDIAN L. INST. (2013).

any infringing copies of the work."

Since NITI Aayog would make its software available to start-ups, who may commercially deploy the software, one could argue that the aforementioned provision would be attracted and the same would be deemed to be an infringement. However, what this argument does not account for is the transformative nature[60] of the software. The same is, arguably, entitled to its own copyright. An enquiry into the transformative nature of the software would require an analysis of the law as it stands.

Indian Courts through various judgements have laid down that "to claim copyright there must be some substantive variation and not just a trivial variation, not the variation of the type where limited ways of expression are available and the author selects one of them." [61] The standard of transformation, in its essence, is a "modicum of creativity."[62] The Supreme Court has held that "the creativity in a derivative work in which the final position will depend upon the amount and value of the corrections and improvements, the independent skill & labour, and the creativity in the end-product is such as to create a new copyright work to make the creato r of the derivative work the author of it".[63]

The High Court of Delhi has also established a *de minimis* threshold for not considering violations as fair dealing within the Act. This threshold essentially means a minor violation, where the Court does not consider trifles. In *India TV Independent News Service Pvt. Ltd. v. Yashraj Films Pvt. Ltd.*, a bench consisting of Justices Nandrajog and Singh laid down the "way forward".[64]

> *"51. The Rule of Law loses its meaning if it does not run close to the Rule of Life. Trivial prima facie violations of copyright are commonplace ...*
>
> *53. It is not in society's best interest to adjudicate these copyright disputes because ultimate compensation paid would not justify public expenditure in the adjudicatory process.*

[60] RG Anand v. Delux Films, AIR 1978 SC 1613 (India).
[61] Syndicate of the Press of the University of Cambridge and v. B.D. Bhandari RFA (OS) No.21 of 2009 and FAO (OS) No.458 of 2008 (High Court of Delhi).
[62] Eastern Book Company v. DB Modak, (2008) 1 SCC 1, 31 (India).
[63] *Id.* at 14.
[64] India TV Independent News Service Pvt. Ltd. v. Yashraj Films Pvt. Ltd. (2012) 192 DLT 502 (India).

54. Secondly, new technologies are emerging which increase the importance of amateur creative production and mix and match creativity. Today amateurs produce creative works of the highest professional quality. Creativity has to be encouraged and this would be in the interest of the society.

55. In our opinion, the use of de minimis, as applied in other areas of the law, without any modification or without having any marriage of convenience, has three significant advantages in the field of Copyright Law. Firstly, the Fair Use concept would be a bad theoretical fit for trivial violations. Secondly, de minimis analysis is much easier. Thirdly, a de minimis determination, is the least time consuming, and needless to state it is in the interest of the parties as also the society that litigation reaches its destination in the shortest possible time.

56. After all, the factors commonly considered by Courts in applying de minimis are well listed. They are five in number: (i) the size and type of the harm, (ii) the cost of adjudication, (iii) the purpose of the violated legal obligation, (iv) the effect on the legal rights of third parties, and (v) the intent of the wrongdoer."

Using Bharat NLP, start-ups who are entrepreneurial amateurs must be encouraged to bring in creativity in technology for the greater interest of society. When tested against the touchstone of the factors for a *de minimis* standard, we see that when a start-up uses the Bharat NLP for a socially relevant purpose, but at the same time making profits, the size and type of harm is negligible and since NITI through this software or end-point would *only* research the nature, context and rules of various Indic languages, the output would be a complete transformation from the data it acquires.

This is akin to a human reading several books and articles in a particular Indic language to further his understanding of the language and ultimately publish and sell a book on an unrelated topic. This would allow the software to further process Indic language contextually and could be licenced to start-ups which use it for various purposes. The cost of adjudication would be exceptionally high due to the vast corpora of data and there is a possibility of numerous trifles adding to the already overburdened adjudicatory system in India. Also, the purpose of the law will not be vitiated since there is no reproduction or substantive reproduction of the literary work. Additionally, it will not have an impact on the rights of third parties due to the transformative nature of the programme and its deployment, and the intent here is not to plagiarise or infringe upon the copyright but to learn the contextual nature of various Indic languages.

Furthermore, it will not amount to "affect prejudicially the owner of the copyright". The framer s must also keep in mind the intersection between technology (NLP), and the regime of copyright law. In India, this intellectual property right does not protect ideas or disembodied information.[65] It is submitted that informational analysis or data mining for the stated technological purpose shall not affect the rights of the copyright holder. The technology merely has insight and does not reproduce the work to the prejudice of the copyright holder. It is important to note that the Courts have devised what would constitute prejudicial. NITI must also reiterate this judicial stance regarding 'unfairness' or 'prejudice' when devising a policy or recommending an amendment to the Copyright Act, 1957. Where the High Court of Lahore, in 1934, held that:

> *"in fair dealing (1) that in order to constitute unfairness there must be an intention to compete and to derive profit from such competition and (2) that unless the motive of the infringer were unfair in the sense of being improper the dealing would be fair."*[66]

In cases where this end-point is used by start-ups, there would be no intention to "compete and to derive profit from such competition" since the outputs are non-substitutable. Where on the one hand, there is a literary work, which provides Bharat NLP to process the text to understand the context of an Indic language. On the other hand, hypothetically, there is a software developed by a start-up which provides information on reproductive health and rights in various Indic languages. There is no malafide intention of the start-up to compete with the literary work. Furthermore, it cannot compete with the work because of the sheer transformed nature of the data.

Therefore, the argument of deemed infringement would also not hold well in court, further NITI may even claim its own copyright over the NLP end-point it develops.

D. LIMITATIONS TO THE LAW

Though the aforementioned argument of fair use may be used for ensuring that the Bharat NLP is available for use by NITI, the law by itself is not robust enough to support the fast-growing nature of technology. This can be seen in the fact that the legislation itself does not explicitly lay

[65] Eastern Book Company, *supra* note 61.
[66] Kartar Singh Giani v. Ladha Singh, AIR 1934 Lah. 777 (High Court of Lahore).

down 'informational analysis' or 'text and data mining'[67] as part of "research" within the exception contained in Section 52 of the Act.

One must keep in mind that these methods do "not use the content itself of the data it analysed. Rather, it relies on the information contained in those documents to be able to draw patterns, conclusions or trends."[68] Further, there are no judicial decisions to analyse the role of AI, Machine Learning, and Natural Language Processing. There, moreover, remains ambiguity regarding liability (if, any at all) when such informational analysis is used for economic gains. That gap in the law needs to be bridged through various changes to the legal architecture.

V. POLICY RECOMMENDATIONS AND CONSIDERATIONS

NITI Aayog should develop its policy on a two-pronged approach. The first, which deals with short term considerations such as the smooth rollout of the Bharat NLP pro gramme. The second, which deals with long- term implications vis-à-vis a liberal framework for AI, NLP and ML.

A. SHORT TERM BASIS

Where NITI can acquire datasets, which are in the public domain, it shall not be infringing on any entity's copyright. Therefore, the first recommendation is that it collates information on which all works are in the public domain. Once this information is collated, it may use the same without any fetters to develop Bharat NLP. This data unlike other copyrighted data may be provided freely to start-ups to possibly train their own systems.

Second, given that there is a strong case that such analytics through the ML and NLP would constitute "fair dealing", the NITI Aayog may proceed with acquiring such data in processing Bharat NLP. That being said, there are the aforementioned ambiguities in the law, and NITI Aayog must prevent any legal challenges to this programme and broader road -blocks to the roll-out of the Digital India programme.

Therefore, it should, *first*, call for the Licensing Policies from all stakeholders since many of the texts they hold will be protected by copyrights. In calling for public data held by governmental or statutory bodies, some may have reserved some rights unto themselves which may act as an impediment to the success of the Bharat NLP programme. Thus making it important to assess

---

[67] Parliamentary Standing Committee on Industry, Science and Technology, *Evidence: Submission by Element AI*, HOUSE OF COMMONS OF CANADA (2018), available at: https://www.ourcommons.ca/DocumentViewer/ en/ 42-1/INDU/meeting-130/evidence.

[68] *Id.*

the Licencing Policies of these stakeholders to begin with; and subsequently, *second*, license the data to create the Bharat NLP. This would ensure that the rights and obligations under the Act as it stands are met, and there is an avoidance of any unnecessary litigation.

B. LONG TERM BASIS

On a long-term basis, NITI Aayog should *first*, conduct a detailed study of legislation and judicial decisions from various jurisdictions and nations on exceptions created for 'informational analysis'. The eco-system must be made conducive for many start-ups and other technology companies to utilise upcoming technologies for the betterment of society at large.

There are two broad approaches, *first*, it is restricted towards certain kinds of use. For instance, the European Union creates a mandatory exception for data mining for certain purposes pertaining to scientific research, education and culture. This is also for a non-profit basis, and only by organisations working for the public interest.[69] In furtherance to this, there is an optional exception which may permit any person to mine data as long as the user has access which is lawful to such data, and the owner of the copyright owner of the data has no reservations for the same.[70]

This, arguably, allows for commercial usage of such data. The United Kingdom also has an exception to informational analysis.[71] This, much like the European Union, is only restricted towards the purposes of research. *Secondly*, Japan has a broad definition of data mining and liberalised usage, even for commercial proposes.[72]

In coming up with a policy, NITI must keep in mind the larger goals of Digital India, 2014,[73] Policy on Adoption of Open Source Software for Government of India, 2012,[74] and National Intellectual Property Rights Policy, 2016.[75]

---

[69] European Parliament, Amended EU Directive on art. 31 (Amendment 8) https://www.europarl.europa.eu/sides-/getDoc.do?pubRef=-//EP//NONSGML+TA+P8-TA-2018-0337+0+DOC+PDF+V0//EN.

[70] *Id.*

[71] Copyright, Designs and Patents Act, 1988, section 29A (Eng.).

[72] Art. 47, Copyright Act of Japan, COPYRIGHT RESEARCH AND INFORMATION CENTER http://www.cric.or.jp/english/clj/cl2.html#cl2_1+A47septies.

[73] Ministry of Electronics and Information Technology, *Digital India*, GOVERNMENT OF INDIA (2014) https://meity.gov.in/sites/upload_files/dit/files/Digital%20India.pdf.

[74] Ministry of Electronics and Information Technology, *Policy on Adoption of Open Source Software for Government of India*, GOVERNMENT OF INDIA (2012) https://meity.gov.in/writereaddata/files/policy_on_adoption_of_oss.pdf.

[75] Ministry of Commerce and Industry, *National Intellectual Property Rights Policy*, GOVERNMENT OF INDIA (2016) https://dipp.gov.in/sites/default/files/National_IPR_Policy_English.pdf. (This policy states its goals, which includes to, "12. Promote use of Free and Open Source Software along with adoption of open standards; possibility

NITI must also keep in mind the evolved jurisprudence of the United States of America where a distinction has been made between expressive and non-expressive use of copyrighted works, in understanding "fair use". Fair use in the US, unlike as interpreted in India, is non-exhaustive and grounded in the principles of fairness and equity. A broader understanding of what may be fair will benefit start-ups, Micro, Small and Medium Enterprises and innovative individuals who may use such AI for the larger benefit of society.

VI. CONCLUSION

A young and ambitious India, on the eve of its seventy-fifth year of existence, faces multiple challenges. A core challenge is the redistribution of resources and access to opportunities. Progress in ICT has demonstrated its usefulness in key development areas like health, education, and governance. To unlock the true potential ICT and its outcomes may hold, it is imperative to improve the access to and through ICT. With the diversity in India, one way to improve this access is to ensure Indic languages are relied on to ensure robust technological development. The access to vast amounts of data in these languages can act as a fillip for new and comprehensive technological development. Ultimately, the purpose of the State should be to seek the larger benefit of the people, which lies in creating an exception to the Copyright Act, 1957, or defining informational analysis within the ambit of "research" within the Act. This would enable the legal framework to keep pace with various technologies which can benefit all persons of India by increasing accessibility to these technologies and various rights made available of facilitated therefrom.

---

of creating Indian standard operating environments will be examined…Legal, technological, economic and soc io-cultural issues arise in different fields of IP which intersect with each other and need to be addressed and resolved by consensus in the best public interest… The present IP Policy aims to integrate IP as a policy and strategic tool in national development plans.").