The Role of Large Language Models in the Recognition of Territorial Sovereignty: An Analysis of the Construction of Legitimacy

FRANCISCO CASTILLO ESLAVA*, University of Granada, Spain
CARLOS MOUGAN*, University of Southampton, United Kingdom
ALEJANDRO ROMERO-RECHE, University of Granada, Spain
STEFFEN STAAB, University of Southampton, United Kingdom and University of Stuttgart, Germany

We examine the potential impact of Large Language Models (LLM) on the recognition of territorial sovereignty and its legitimization. We argue that while technology tools, such as Google Maps and Large Language Models (LLM) like OpenAI's ChatGPT, are often perceived as impartial and objective, this perception is flawed, as AI algorithms reflect the biases of their designers or the data they are built on. We also stress the importance of evaluating the actions and decisions of AI and multinational companies that offer them, which play a crucial role in aspects such as legitimizing and establishing ideas in the collective imagination. Our paper highlights the case of three controversial territories: Crimea, West Bank and Transnitria, by comparing the responses of ChatGPT against Wikipedia information and United Nations resolutions. We contend that the emergence of AI-based tools like LLMs is leading to a new scenario in which emerging technology consolidates power and influences our understanding of reality. Therefore, it is crucial to monitor and analyze the role of AI in the construction of legitimacy and the recognition of territorial sovereignty.

CCS Concepts: • Social and professional topics \rightarrow Political speech; • Human-centered computing \rightarrow Empirical studies in collaborative and social computing; Empirical studies in HCI; Field studies.

Additional Key Words and Phrases: Geopolitics, Generative Models, Large Language Models, Social Sciences

1 INTRODUCTION

Territorial sovereignty is a topic that is not typically free of controversy [Peter 2017]. When more than one political actor demands the right to administer the same region, it forces the remaining actors to take a position on the new dilemma that opens [Weber 2009]. Recognizing the sovereignty of one of the parties involved carries the possibility of animosity with the aggrieved party and granting legitimacy with respect to the disputed territory. In the case of states, they have their own tools, such as diplomatic recognition, to vest a political subject with authority over a territory [Weber 2016].

But it is not only states that confer legitimacy to different political authorities. Organizations and multinational corporations play a crucial role in legitimizing and establishing ideas in the collective imagination [Habermas 1985]. This means that today, borders are not defined solely by formal political institutions. Considering that authority has a right over a territory or the perception that a certain entity administers a region is also determined by the position that companies take and reflect on the issue. Consequently, it is crucial to closely monitor and evaluate the actions and decisions taken by these organizations.

A clear example, is Alphabet Inc. through its well-known online mapping service Google Maps. The difficulty of delimiting territory and the inherent arbitrariness that characterizes the task of drawing borders in the presence of a territorial conflict led to the offense and complaints of more

Authors' addresses: Francisco Castillo Eslava*, University of Granada, Spain, fceslava@ugr.e; Carlos Mougan*, University of Southampton, United Kingdom, c.mougan-navarro@soton.ac.uk; Alejandro Romero-Reche, University of Granada, Spain, romeroreche@ugr.es; Steffen Staab, University of Southampton, United Kingdom and University of Stuttgart, Germany, S.R.Staab@soton.ac.uk.

than one political actor who felt damaged by the company's defined border demarcation. As reflected in the work of Bogen and Quiquivix [Miranda 2016; Quiquivix 2014], the delimitation of the region that houses the Palestinian-Israeli conflict or the territory of Arunachal Pradesh claimed by China and India are two examples of the different controversies generated around the application.

In recent years, driven precisely by the investment of these types of technology companies, a set of artificial intelligence-based tools has been developed. Its application and use have become popular due to its performance. One aspect that characterizes them, besides the speed, efficiency, and scalability with which they operate, is the impartiality often attributed to them. This perceived impartiality comes from considering that their verdicts are the result of operating with objective data provided without the mediation of a third party[Fountain 2022]. The main problem and threat that arise from the social legitimization of this consideration are that it is actually a perception. The basis on which these types of applications are built is the vast amount of information extracted, mostly, from the Internet. This means that the results offered by the tool will be exposed to the same biases that are found in the data, or alternatively, will be flawed by the designer's will[Joyce et al. 2021].

One of the latest online applications that has attracted the most media attention is being carried out by OpenAI's ChatGPT. These new algorithms called Large Language Models (LLM) are oriented towards interaction with the user through questions and answers, keeping a high degree of naturalness in the development of the conversation. The enormous impact generated by its appearance outlines a new scenario in which this type of emerging technology ends up consolidating itself as a new information search tool. The consolidation of this process would integrate them into the social field as a new actor legitimizing public opinion and knowledge. In this way, the imminent influence that the application and popularity of this technology can reach in society makes its study a necessity and impels us to take an interdisciplinary perspective capable of addressing the different dimensions that characterize it.

Thus, the objective of our research is to examine the positions that ChatGPT reflects on issues that constitute a public debate. Existing territorial conflicts of sovereignty will be used for a case study due to their complexity, importance, and implications, both sociol and in the territorial conformation of the states.

In summary, our research makes the following contributions:

- First, it highlights the impact of new technologies on the construction of legitimacy.
- Second, it takes ChatGPT as a new legitimizing social actor of knowledge and provides an analysis and comparison of its responses on the legitimacy of three disputed territories with the position reflected by the United Nations and Wikipedia.
- Third, it contributes to the ongoing discussion about the risks and potential harm arising from the usage of Large Language Models

2 BACKGROUND

2.1 Large Language Models

Large Language Models (LLMs) are machine learning models that are designed to represent the likelihood of word sequences in a given domain, such as essays, books, or documents. LLMs are trained to capture statistical patterns in word sequences from a particular corpus, and can then be used to predict the probability of generating the next sequence of words. The architecture and training objectives of an LLM may vary depending on its intended use. In this discussion, we will focus specifically on LLMs that are tailored for language generation.

A series of Large Language Models (LLMs) [Brown et al. 2020; Lieber et al. 2021; Rae et al. 2021; Thoppilan et al. 2022], with the largest dense autoregressive transformers [Vaswani et al. 2017]

Table 1. Generative text model outputs temporal evolution comparison. 2011 is extracted from Sutskever et al. 2011], 2020 text from GPT-3 [Goldstein et al. 2023a; Vaswani et al. 2017], and 2023 by querying chatGPT (6 March 2023)

2011

The meaning of life is the tradition of the ancient human reproduction: it is less favorable to the good boy for when to remove her bigger

2020

The meaning of life is a philosophical and existential question that has been debated for centuries. There is no single answer or definitive explanation for what the meaning of life is, as it can vary from person to person based on their beliefs, values, and experiences. Some people believe that the meaning of life is to achieve happiness and fulfillment, while others believe it is to seek knowledge and understanding of the world around us. Some believe that the purpose of life is to fulfill a certain destiny or to contribute to society in a meaningful way.

Others see life as inherently meaningless, and that it is up to each individual to create their own meaning and purpose through their choices and actions.

2023

The meaning of life is a philosophical and existential question that has been debated for centuries. There is no single answer or definitive explanation for what the meaning of life is, as it can vary from person to person based on their beliefs, values, and experiences. Some people believe that the meaning of life is to achieve happiness and fulfillment, while others believe it is to seek knowledge and understanding of the world around us. Some believe that the purpose of life is to fulfill a certain destiny or to contribute to society in a meaningful way. Others see life as inherently meaningless, and that it is up to each individual to create their own meaning and purpose through their choices and actions. Ultimately, the meaning of life is a deeply personal and subjective concept that each person must explore and define for themselves.

having over 500 billion parameters, have been introduced. These LLMs have demonstrated impressive performance on various tasks, using evaluation protocols such as zero-shot, few-shot, and fine-tuning.

Some of these methods, including ChatGPT, have been developed using a technique known as Reinforcement Learning from Human Feedback [Christiano et al. 2017]

. While this language model training technique can create highly conversational algorithms, it is still in a development phase, and several examples have shown blatant errors, such as referencing non-existent scientific studies [Thorp 2023]. As a result, the impact of these algorithms on society is yet to be fully realized.

Fine-tuning with crowd worker-annotated data, which is the last step in the training process, can result in significant improvements in ensuring that the model's responses are consistent with a set of human values, such as preventing harmful suggestions and unfair bias [Thoppilan et al. 2022]. [Solaiman and Dennison 2021] show that adjusting LLMs behavior is feasible with a small sample of hand-curated dataset. Although current open-source resources for LLMs are trained using this fine-tuning step, there are still additional steps pending, such as factual grounding. This involves enabling the model to consult external knowledge sources, such as information retrieval systems, language translators, and calculators. The aim of factual grounding is to enable the model to generate responses based on known sources, rather than responses that merely seem plausible. This approach may be useful in the future for improving the reliability of LLM-generated responses.

2.2 Social Science Foundations

Responsible innovation requires an interdisciplinary approach [Stilgoe et al. 2013]. In order to anticipate, understand and mitigate the possible risks posed by new technology, there is the need to view its implications through different lenses. In this work we provide a territorial construction legitimacy perspective.

As conceptualized by social theory and established by empirical social research, all social interaction takes place within the framework of specific structural constraints and cultural patterns and, in its turn, produces, reproduces, or changes such framework for future interactions. Human interaction with AI is bound to develop a similar dynamic [Airoldi 2022]: both humans and algorithms learn from each interaction and adjust for future encounters in a feedback loop that reinforces certain patterns while discouraging others.

However fluid or convincing the use of natural language, interacting with ChatGPT or any other AI is essentially different, at a sociological level, from interacting with another human being, at least to the extent that the human user is aware of the non-human nature of their partner in interaction. The social implications of such nature depend on how the non-human agent has been culturally constructed and the qualities it is perceived as having. Users looking for answers about relevant topics will value them differently if the non-human agent is perceived as being impartial or biased, comprehensive or arbitrary, etc.

Obviously, this is particularly relevant when AI is used not just as a source of unstructured factual information but, rather, as an authoritative source of structured knowledge, organized in pondered, essay-like responses where conflicting approaches and positions seem to be carefully weighed. In such cases, the responses provided may be perceived by users, due to the non-human nature of the agent, as fair and impartial in a way unattainable to human agents, even if the process of information-gathering is actually shaped by intrinsic biases in the original sources used or in their selection.

Hence our focus in the recognition of territorial sovereignty and the potential role of ChatGPT in the social perception of legitimacy. Sovereignty is socially constructed, both in specific historical instances and in its meaning as a concept [Glanville 2013]; it is crucially shaped by the degree of recognition granted by international actors, and also by how statehood claimants manage such recognition or lack of it [Kyris 2022]. While the analytical approach to sovereignty in International Relations tends to be macroscopic, highlighting the roles played by nations, organizations and institutions, a thorough understanding of the process requires taking social perceptions into account, and how these are influenced by media, pundits, experts and other information and knowledge sources.

As opposed to sources, such as Wikipedia entries, conspicuously written by humans and often suspected as biased despite editorial procedures designed to ensure objectivity, responses produced by AI are not expected to reflect the moral or political values of a writer or a team of writers located in specific countries and, therefore, defending specific interests in a conscious way or not. Algorithmic judgment appears as the way to avoid the essentially suspicious human subjectivity [Carlson 2018], and therefore its influence in social legitimation processes is potentially greater than those of other sources whose biases are taken for granted.

This calls for at least two complementary research approaches: 1) public opinion research about how AI is perceived as a source of knowledge by the general population, and 2) comparative discourse analysis of ChatGPT responses, considering that, while these remain inconclusive about unsettled political or scientific debates, they could fundamentally shape those debates through tacit assumptions.

2.3 Related work

With the continued advancement of generative models, including language models (as shown in Table 1), there is growing concern among researchers, practitioners, and commentators about the potential benefits, risks, and negative effects associated with these models. To address this, workshops have been organized to better understand the risk landscape, specifically regarding the emergence of LLMs [Goldstein et al. 2023b; Tamkin et al. 2021]. Additionally, several reports have been published that provide: a comprehensive analysis of foundation models, covering various aspects such as technical principles, capabilities, applications, and societal impact [Bommasani et al. 2021]; help structure the risk landscape associated with LLMs providing multidisciplinary literature [Weidinger et al. 2021]; emphasizes environmental, financial costs of data curation, collection, and documentations [Bender et al. 2021]; safety problem landscape for end-to-end conversational AI [Dinan et al. 2021]; discuss the impacts of code generation technologies [Chen et al. 2021]; study how generative AI will increase productivity for knowledge workers [Noy and Zhang 2023]. In our work, we focus on a particular critical use case, construction of legitimacy of territorial recognition, that to the best of our knowledge has not been presented within the literature, and we provide an analysis comparing against different sources.

With respect to the evaluation of mitigating toxicity strategies in language models, [Welbl et al. 2021] analyze the consequences of these strategies in terms of model bias and quality, and evaluate their effectiveness using both automatic and human evaluation methods. In our work, we do not focus on mitigating toxicity but on evaluating our position on the topic of territorial recognition in comparison to other agents.

[Kenton et al. 2021] highlight some ways behavioral issues of LLMs can arise from accidental misspecification by the system designer. Our work is not necessarily targeted to mispecifications, but to highlight the difficulty of the task, which even among politics experts is not typically free of controversy [Peter 2017].

The authors of the paper titled "Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets" [Solaiman and Dennison 2021] propose an iterative process that aims to change the behavior of language models significantly. This process involves crafting and fine-tuning the language model on a dataset that reflects a predetermined set of target values. The authors demonstrate that it is feasible to adjust language model behavior significantly with a small, hand-curated dataset. In our work, we analyze how generative language models can provide knowledge about the recognition of territorial sovereignty, which opens up the question of whether it's possible to tackle the problem of LLM-human alignment even in cases when there is no human consensus and how to react in these situations.

An example of the effects of technology on territorial sovereignty is Alphabet Inc.'s online mapping service Google Maps, which has led to complaints from political actors who felt damaged by the company's defined border demarcation. The delimitation of the region that houses the Palestinian-Israeli conflict or the territory of Arunachal Pradesh claimed by China and India are two examples of the different controversies generated around the application [Miranda 2016; Quiquivix 2014].

3 METHODOLOGY

The work was carried out through a content analysis of the answers provided by ChatGPT, the information found on Wikipedia and the resolutions issued by the General Assembly and the United Nations Security Council. Their application made it possible to define the position of each anonymized actor. Wikipedia's position was added in order to be able to compare the significance

of ChatGPT with another widely spread information search web resource. And, the UN position was taken as an expression of the international community to check its possible influence.

Three territories were selected for analysis: Crimea, the West Bank, and Transnistria. The first was chosen because it represents one of the most relevant territorial conflicts of the last decade. The case of the West Bank was chosen because of the historical dimension that has characterized the conflict and the difficulties generated by the diplomatic relations that Israel maintains with a large part of the international community. And Transnistria was added because of the repeated occasions on which the UN Security Council has urged respect for the sovereignty and territorial integrity of the Republic of Moldova.

In the case of the IA, the data collection was obtained on December 28, 2022 through a series of questions asked to the three territories:

- Question 1: What is X?
- Question 2: To whom should X belong?
- Question 3: Who has rights over X?

The construction of the questions set out was based on two approaches to the model: a general one that would serve as a contextualization of the territory, and another more direct approach to what the research was intended to find out. The first question was intended to find out if the model assigns to any political actor the sovereignty of the territory about which it is asked. With the statement of the other two questions, the model was already made aware of the existence of a dispute over the territory being asked about and it was desired to check the degree to which ChatGPT would be signified by exposing a disjunctive. It will be considered that the model reflects a neutral attitude when it does not take the side of any of the parties to the conflict and, arbitrary, when the impartiality of the training data conditions it in favor of one of the parties.

4 EXPERIMENTS

Table 2. Results table with a summary of the position of the different actors analyzed towards the recognition of territorial sovereignty

Territory	Recognition		
	United Nations	Wikipedia	ChatGPT
Crimea		In favor of Ukraine	
West Bank	In favor of Palestine	In favor of Palestine	In favor of Palestine
Transnistria	In favor of Moldavia	In favor of Moldavia	Neutral

4.1 Crimea

4.1.1 United Nations. The position of the United Nations as an organization regarding the sovereignty of the Crimean territory was last resolved on the side of Ukraine in resolution 68/262 on March 27, 2014 [UN/A/RES/68/262 2014]. Adopted with 100 votes in favor, 11 against, and 58 abstentions, the General Assembly:

"Affirms its commitment to the sovereignty, political independence, unity and territorial integrity of Ukraine within its internationally recognized borders" "Underscores that the referendum held in the Autonomous Republic of Crimea and the city of Sevastopol on 16 March 2014, having no validity, cannot form the basis for any alteration of the status of the Autonomous Republic of Crimea or of the city of

Sevastopol"

"Calls upon all States, international organizations and specialized agencies not to recognize any alteration of the status of the Autonomous Republic of Crimea [...] on the basis of the above-mentioned referendum and to refrain from any action or dealing that might be interpreted as recognizing any such altered status"

- 4.1.2 Wikipedia. The online encyclopedia describes the situation as a clear occupation by Russia and considers the territory a region of Ukraine from the very first sentence: "Crimea is a peninsula in Ukraine, on the northern coast of the Black Sea, that has been occupied by Russia since 2014". Still in the introductory section, the website again speaks of Russian occupation, calls the referendum and annexation illegal and informs the reader that most of the international community recognizes Ukraine's sovereignty over the territory: "In 2014, the Russians occupied the peninsula and organized an illegal referendum in support of Russian annexation, but most countries recognize Crimea as Ukrainian territory". This position is reaffirmed and the same arguments are restated in the history subsection under "Russian occupation (2014-present)" [Wikipedia 2023a].
- 4.1.3 ChatGPT. From ChatGPT's answers a declination towards the Ukrainian part of the conflict is induced. In the second sentence of the first question it already means considering the territory part of Ukraine, although it warns of being controlled by Russia: "It is located in Eastern Europe and is part of Ukraine, although it is currently controlled by Russia".

The answers to the next two questions (Q2 and Q3), retain practically the same content and much of the same structure. In both, the model adds that the annexation has not been recognized by most countries of the international community, states that it has been widely condemned as a violation of international law, and refers to the resolutions in which the United Nations General Assembly urges Russia to withdraw its forces from Crimea and to respect the sovereignty and territorial integrity of Ukraine.

The difference between the latter two parts comes from the degree to which they ratify Ukraine's sovereignty over Crimea. While the second question displays a more descriptive tone of the conflict and describes the sovereignty of the territory as an object in international dispute - "the sovereignty of Crimea is a matter of international dispute" -, the third one leaves no room for doubt as to who has sovereignty over the area and turns to international law:

"'According to international law, Ukraine has the right to sovereignty over Crimea. Crimea is an autonomous republic within Ukraine and has been recognized as such by most countries in the international community'"

Another aspect by which ChatGPT's position on the conflict can be perceived stems from the total absence of any statement, argument or version that could legitimize the action taken by Russia.

4.2 West Bank

- 4.2.1 United Nations. For the territory comprising the Palestinian-Israeli conflict, the position of the United Nations has been expressed through the issuance of a series of Security Council resolutions in favor of Palestine: 242 (1967), 338 (1973), 446 (1979), 452 (1979), 465 (1980), 476 (1980), 478 (1980), 1397 (2002), 1515 (2003), 1850 (2008) and 2334 (2016). If we take as a reference the last resolution [UN/S/RES/2334 2016], although they all coincide in the sense of the ruling, we can infer where the organization establishes the territorial limits of the two States when it expresses its concern about the fact that the:
- "Continuing Israeli settlement activities are dangerously imperiling the viability of the two-State solution based on the 1967 lines..."

The 1967 lines refer to the territorial demarcation established in the Arab-Israeli armistice of 1949. This would come to be adopted as the reference limits with which to recognize Palestinian

territory after Israel exceeded them in the June 1967 war. In consideration of the 1967 lines, the West Bank is considered part of the borders of the Palestinian State.

Returning to the text, the UN condemns any measures undertaken by Israel "aimed at altering the demographic composition, character and status of the Palestinian Territory occupied since 1967" and refers to resolution 1515 (2003) to remind it of the obligation to halt "all settlement activity, including 'natural growth', and the dismantlement of all settlement outposts erected since March 2001". Thus, the organization "reaffirms that the establishment by Israel of settlements in the Palestinian territory occupied since 1967, including East Jerusalem, has no legal validity and constitutes a flagrant violation under international law".

- 4.2.2 Wikipedia. For the issue of the West Bank, Wikipedia again refers to the situation as an occupation. In this case, it states that the territory is "under an Israeli military occupation since 1967". Another description that confirms this version comes when it refers to Israelis living in the West Bank as settlers. To reinforce the position it conveys, the website adds as an argument the illegality with which the international community considers Israeli settlements on the basis of international law. In the same way, it uses a ruling of the International Court of Justice as an endorsement. Apart from the introduction, Wikipedia again deals with the subject under the same terms in the different sections, such as in the history section and in the specific section called "Consequences of occupation" [Wikipedia 2023c].
- 4.2.3 ChatGPT. On this occasion, there is not so much evidence to perceive a clear positioning of ChatGPT on the conflict. In fact, it is only discernible from the use of one word in the first general question. Specifically when it uses the term "occupied" when describing the configuration of territory: "Cisjordan includes the modern state of Israel, as well as the West Bank and Gaza Strip, which are territories occupied by Israel but claimed by the Palestinians". With this sentence he delegitimizes Israel's sovereignty over the territories in the West Bank and grants sovereignty to Palestine.

On the other hand, the results obtained from the other two questions become an exercise in contextualizing the conflict. Two almost identical texts are extracted in which an attempt is made to reflect an impartial and merely descriptive position of the conflict:

"The question of who has rights over Cisjordan, also known as Palestine or the Land of Israel, is a complex and controversial issue with a long history"

"Some people believe that the western region of Cisjordan should belong to the state of Israel, while others believe it should be an independent Palestinian state."

"There are different viewpoints on this issue, and the resolution to the conflict will depend on the willingness of the parties involved to find a mutually acceptable solution that takes into account the rights and interests of all parties"

4.3 Transnistria

4.3.1 United Nations. The position of the United Nations on the status of Transnistria is again induced from the set of resolutions issued on the territory, but in this case by the General Assembly: 54/117 (1999), 55/179 (2000), 56/216 (2001), 57/298 (2003)[UN/A/RES/57/298 2003] and 72/282 (2018)[UN/A/RES/72/282 2018]. In the first two, the existing references to the Transdniestrian issue are aimed at supporting efforts to resolve the conflict and to reflect Russia's commitment to withdraw its forces from the territory of the Republic of Moldova. The two successor resolutions to those just cited retain the same purpose, but add that the political settlement of the conflict must be based on "full respect of the sovereignty and territorial integrity of the Republic of Moldova". Finally, the last pronouncement of the General Assembly is issued in order to urge "the Russian"

Federation to complete, unconditionally and without further delay, the orderly withdrawal of the Operational Group of Russian Forces and its armaments from the territory of the Republic of Moldova". In it, the UN recalls that "the stationing of foreign military forces on the territory of the Republic of Moldova, without its consent, violates its sovereignty and territorial integrity" and is concerned that "the continuous illegal joint military exercises of the Operational Group of Russian Forces with the paramilitaries of the separatist entity in the eastern part of the country, [...], disregards the sovereignty and territorial integrity of the Republic of Moldova.

4.3.2 Wikipedia. Transnistria is defined as an "unrecognized breakaway state" and reports that it is internationally recognized as part of Moldova. Despite this, it recognizes that the Pridnestrovie Republic of Moldova controls a large part of the area. Still sticking to the first paragraph, Wikipedia makes clear the scarce international recognition of the region, warning that it only has the support of three unrecognized or partially recognized separatist states - Abkhazia, Artsakh and South Ossetia-. In the same vein and with the intention of manifesting Russian influence and dependence, it mentions the resolution of the Parliamentary Assembly of the Council of Europe defining the territory as under military occupation by Russia. The content of the website intertwines statements that allude to the de facto control of the region as an independent republic with others that recall the scarce international recognition. Finally, in the section dedicated exclusively to the status of the region[Wikipedia 2023b], the web site pronounces that:

"All UN member states consider Transnistria a legal part of the Republic of Moldova. Only the partially recognised or unrecognised states of South Ossetia, Artsakh, and Abkhazia have recognised Transnistria as a sovereign entity after it declared independence from Moldova in 1990 with Tiraspol as its declared capital".

4.3.3 ChatGPT. Finally, in the case of Transnistria, it was the only territory where no evidence was found in the language used by ChatGPT to draw a clear conclusion about its position on the conflict. As was seen in much of the West Bank example, the model's response offers a contextualization of the conflict with great care in the language so as not to be biased. Thus, the response is limited to providing historical data and indicating the position of the actors involved from a neutral position:

"Transnistria declared independence from Moldova in 1990, but its independence has not been recognized by any country"

"The Moldovan government considers Transnistria to be an autonomous territorial unit within Moldova, while the Transnistrian authorities consider themselves to be an independent state"

The only aspect that can be reproached in ChatGPT's response with respect to the neutrality it pursues in the matter is the failure to allude to the recognition granted by the three unrecognized or partially recognized separatist states: Abkhazia, Artsakh and South Ossetia.

5 DISCUSSION

The content analysis undertaken on the ChatGPT responses and comparison against Wikipedia and the United Nations resolutions has led to the identification of a set of particularities that have facilitated the task of characterizing the behavior displayed by the model. The main issue to highlight refers to an analysis of the biased answers of AI systems. As we have been able to verify, the answers offer, in two of the three cases, evidence that reflects a position in favor of one of the parties to the conflict. For Crimea, it explicitly positions itself on the side of Ukraine and, although it is less categorical with regard to the West Bank, it offers a biased response in favor of Palestine when speaking of occupation with respect to the territories controlled by Israel in the region.

The case of Transnistria is the only territory where no terms or expressions have been found to induce the AI to take sides in the conflict. Taking into account that part of ChatGPT's source of information is based on Web content, it is possible that the significantly lower media coverage of the conflict has influenced the amount of information generated in this regard and, thus, the neutrality of the model. On the other hand, the categorical expressions with which he addresses the issue of the Crimean peninsula may be due to precisely the opposite:: one of the most important international conflicts of the last decade and with the condemnation of the Russian annexation by most of the international community -as reflected in the UN General Assembly resolution 68/262 [UN/A/RES/68/262 2014].

In relation to the Palestinian-Israeli issue, a possible explanation for the less forcefulness with which ChatGPT is signified and the descriptive nature of its responses can perhaps be found in the high complexity of the conflict and in the actors involved. Although the UN Security Council has already pronounced itself warning of the illegality of the Israeli settlements in the West Bank [UN/S/RES/2334 2016]-with the abstention of the US-, the close relations that Israel maintains with much of the international community might influence attenuate the positions of its partners. This could significantly affect the generation of the number of official documents condemning Israel's conduct. Just as the use of gender-ethnicity based image ratios has been a topic of controversy in the past [Buolamwini and Gebru 2018], it prompts the question of what official document ratios should be used for international political conflicts and, consequently, the source of data from which the model learns would also be reduced.

Thus, the position of the two online resources has always taken the same position as the one issued by the UN, except for the case of Transnistria in which ChatGPT takes a neutral position on the conflict. The fact induces to think that the attitude of the international community influences the position reflected by ChatGPT and the content added to Wikipedia.

Another aspect worth highlighting is related to the effectiveness of the questions developed to achieve the research objective. Two types of questions were applied to the model: (i) the first of a general nature about the territory and (ii) the other aimed directly at finding out who ChatGPT considers has more rights over the disputed territory.; it was the first one that best captured the positioning of the model. This may be due to the training to which the AI has been subjected to detect questions that expect a biased or tendentious answer from the model. Thus, the most effective strategy in our research to collect the behavior of ChatGPT has been to ask general and indirect questions about the issue to be evaluated, instead of specific and directed directly to decipher the position of the model. This way of skirting the impartiality with which the AI responds offers greater possibilities of reflecting an arbitrary behavior in response to a question asked.

Finally, an issue perceived in the work and which has a bearing on the idea just expressed has to do with the high similarity found between the answers to questions two and three. In these questions, in addition to the content whose difference is minimal when dealing with the same territory, a general pattern is observed that is maintained in the responses from the different regions:

"The question of to whom Transnistria belongs is a complex and disputed issue" (Question 2. Transnistria)

"The question of who has rights over West Bank [...] is a complex and controversial issue with a long history" (Q3. West Bank)

"Ultimately, the resolution of disputes over sovereignty and territorial control depends on the willingness of all parties to engage in dialogue and find mutually acceptable solutions" (Q3. Transnistria) "Ultimately, the resolution to this conflict will depend on the willingness of the parties involved to find a mutually acceptable solution that takes into account the rights and interests of all parties "(Q2. West Bank)

This singularity shows the last step of ChatGPT training process Reinforcement Learning from Human Feedback [Christiano et al. 2017] (cf. Section 2), with respect to the willingness of ChatGPT developers to identify and avoid questions that seek to provoke the arbitrariness of the model. The detection of patterns built by similar expressions in similar questions leads us to appreciate the semantic structure applied during the training process. The detection of these patterns in the semantics of the generative model reduces at the same time the naturalness with which it interacts.

6 CONCLUSION

The results and the analysis carried out throughout the study have led to a series of reflections on the behavior identified in ChatGPT. The first one is directed to the behavior exhibited by the model when expressing itself on a controversial topic. At this juncture, artificial intelligence LLMs aspire to offer a neutral response that does not lean towards any of the opposing parties. Despite attempts to avoid bias, the results of the work have shown that in two of the three scenarios the model has ended up offering a biased response. This perceived attitude has significant implications for both the knowledge acquisition processes and the case study selected for this research.

Algorithmic judgment, as performed by ChatGPT when questioned about territorial sovereignty, may avoid certain elements of human subjectivity, but still incorporates value judgments and positions that may be intrinsic to the sources from which it gathers information or even to the very definition of concepts and problems. The emergence of a computational authority [Airoldi, 2021], regarded as a key source of knowledge that impartially ponders the merits and arguments of conflicting positions to provide an unbiased review of the subject matter, could entail the reinforcement of specific points of view, thus established as objective by a non-human, dispassionate (and, therefore, more reliable) intelligence.

To the extent that LLMs are popularized to a mass public as an information search resource, their impact on the generation and consolidation of knowledge will be extended to all areas susceptible to their use. In the case at hand, the potential implications are especially significant because they transcend the individual level and also have an effect on state borders. The recognition of a territory to a political actor constitutes the legitimization of the one who issues it. With this in mind, the perception of all users who make use of AI such as ChatGPT on the sovereignty of a region will be influenced by the positioning of the model's responses with respect to the territory. Thus, the analyzed technology becomes a relevant social agent in the consolidation of public opinion about the territorial sovereignty of a region, especially in those territories where more than one political actor claims sovereignty over it. It should be added that these resources are being offered by large technology companies whose orientation may be conditioned to the particular interests of the companies.

In view of the above and the biased results that the model has offered, it becomes necessary to address the debate on the role that these new tools should play in society and the position that they should reflect in the face of the transcendental dilemmas described above. Opting for neutrality by offering a fair description of the conflict or reproducing the position taken by the international community are two options that are not free of risks and controversies. The issue is extremely complex and multifaceted, and ongoing debates are needed to understand the impact and consequences of AI behavior on knowledge generation and territorial legitimization. Thus, with the development of this paper we hope to make visible one of the conflicts that generates the

greatest tension at the international level and in which artificial intelligence has begun to play a leading role.

Limitations and further work: This work opens up a problem and necessarily leaves questions unanswered. For example, we do not discuss potential beneficial applications of LMs nor do we focus in many international conflicts nor on different languages. Even though territorial recognition problems are a longstanding political issue, the technology used represents a snapshot in time, and algorithms are queried through the years 2022 and 2023.

REFERENCES

Massimo Airoldi. 2022. Machine habitus: Toward a sociology of algorithms. John Wiley & Sons, .

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 FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021, Madeleine Clare Elish, William Isaac, and Richard S. Zemel (Eds.). ACM, ., 610–623. https://doi.org/10.1145/3442188.3445922

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya 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 D. 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 S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. 2021. On the Opportunities and Risks of Foundation Models. *CoRR* abs/2108.07258 (2021), arXiv:2108.07258 https://arxiv.org/abs/2108.07258

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). , , , https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA (Proceedings of Machine Learning Research, Vol. 81), Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html

Matt Carlson. 2018. Automating judgment? Algorithmic judgment, news knowledge, and journalistic professionalism. New media & society 20, 5 (2018), 1755–1772.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating Large Language Models Trained on Code. CoRR abs/2107.03374 (2021), . arXiv:2107.03374 https://arxiv.org/abs/2107.03374

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences. In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). ., ., 4299– 4307. https://proceedings.neurips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html

Emily Dinan, Gavin Abercrombie, A. Stevie Bergman, Shannon L. Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2021. Anticipating Safety Issues in E2E Conversational AI: Framework and Tooling. *CoRR* abs/2107.03451 (2021), arXiv:2107.03451 https://arxiv.org/abs/2107.03451

- Jane E Fountain. 2022. The moon, the ghetto and artificial intelligence: Reducing systemic racism in computational algorithms. Government Information Quarterly 39, 2 (2022), 101645.
- Luke Glanville. 2013. The myth of "traditional" sovereignty. International Studies Quarterly 57, 1 (2013), 79-90.
- Josh A. Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023a. Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. CoRR abs/2301.04246 (2023), ,. https://doi.org/10.48550/arXiv.2301.04246 arXiv:2301.04246
- Josh A. Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023b. Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. https://doi.org/10.48550/ARXIV.2301.04246
- Jürgen Habermas. 1985. The theory of communicative action: Volume 1: Reason and the rationalization of society. Vol. 1. Beacon press, ,.
- Kelly Joyce, Laurel Smith-Doerr, Sharla Alegria, Susan Bell, Taylor Cruz, Steve G Hoffman, Safiya Umoja Noble, and Benjamin Shestakofsky. 2021. Toward a sociology of artificial intelligence: A call for research on inequalities and structural change. Socius 7 (2021), 2378023121999581.
- Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. 2021. Alignment of Language Agents. CoRR abs/2103.14659 (2021), . arXiv:2103.14659 https://arxiv.org/abs/2103.14659
- George Kyris. 2022. State recognition and dynamic sovereignty. European journal of international relations 28, 2 (2022), 287–311.
- Opher Lieber, Or Sharir, Barak Lenz, and Yoav Shoham. 2021. Jurassic-1: Technical details and evaluation. White Paper. AI21 Labs 1 (2021), .
- Bogen Miranda. 2016. Is Google wired for geopolitics? https://medium.com/@mbogen/is-google-wired-for-geopolitics-77ae9a2d7fa2 [Online; posted 14-August-2016].
- Shakked Noy and Whitney Zhang. 2023. Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence. . . , . (2023), .
- Fabienne Peter. 2017. Political Legitimacy. In *The Stanford Encyclopedia of Philosophy* (Summer 2017 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University, .
- Linda Quiquivix. 2014. Art of War, Art of Resistance: Palestinian Counter-Cartography on Google Earth. *Annals of the Association of American Geographers* 104, 3 (2014), 444–459. https://doi.org/10.1080/00045608.2014.892328 arXiv:https://doi.org/10.1080/00045608.2014.892328
- Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew J. Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William S. Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. CoRR abs/2112.11446 (2021), arXiv:2112.11446 https://arxiv.org/abs/2112.11446
- Irene Solaiman and Christy Dennison. 2021. Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets. In Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual, Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.). NeurIPS 2021, ., 5861–5873. https://proceedings.neurips.cc/paper/2021/hash/2e855f9489df0712b4bd8ea9e2848c5a-Abstract.html
- Jack Stilgoe, Richard Owen, and Phil Macnaghten. 2013. Developing a framework for responsible innovation. Research policy 42, 9 (2013), 1568–1580.
- Ilya Sutskever, James Martens, and Geoffrey E. Hinton. 2011. Generating Text with Recurrent Neural Networks. In Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 July 2, 2011, Lise Getoor and Tobias Scheffer (Eds.). Omnipress, ., 1017–1024. https://icml.cc/2011/papers/524_icmlpaper.pdf
- Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. 2021. Understanding the capabilities, limitations, and societal impact of large language models. arXiv preprint arXiv:2102.02503., . (2021), .
- Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts,

Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed H. Chi, and Quoc Le. 2022. LaMDA: Language Models for Dialog Applications. *CoRR* abs/2201.08239 (2022), arXiv:2201.08239 https://arxiv.org/abs/2201.08239 H Holden Thorp. 2023. ChatGPT is fun, but not an author., 313–313 pages.

Jnited Nations: UN/A/RES/57/298. 2003. General Assembly:Cooperación entre las Naciones Unidas y la Organización para la Seguridad y la Cooperación en Europa. , 7 pages. https://documents-dds-ny.un.org/doc/UNDOC/GEN/N02/561/20/PDF/N0256120.pdf?OpenElement

United Nations: UN/A/RES/68/262. 2014. General Assembly: Resolution adopted by the General Assembly on 27 March 2014. , 2 pages. https://documents-dds-ny.un.org/doc/UNDOC/GEN/N13/455/17/PDF/N1345517.pdf?OpenElement

United Nations: UN/A/RES/72/282. 2018. General Assembly:Resolución aprobada por la Asamblea General el 22 de junio de 2018., 3 pages. https://documents-dds-ny.un.org/doc/UNDOC/GEN/N18/195/42/PDF/N1819542.pdf?OpenElement United Nations: UN/S/RES/2334. 2016. Security Council:Resolution 2334 (2016). , 3 pages. https://www.un.org/webcast/pdfs/SRES2334-2016.pdf

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 Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). Neural Information Processing Systems 2017, ., 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Max Weber. 2009. The theory of social and economic organization. Simon and Schuster, .

Max Weber. 2016. The types of legitimate domination. In Social Theory Re-Wired. Routledge, ., 270-286.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359. (2021),.

Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. 2021. Challenges in Detoxifying Language Models. In Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, "2447–2469. https://doi.org/10.18653/v1/2021.findings-emnlp.210

 $Wikipedia.\ 2023a.\ Crimea-Wikipedia, The\ Free\ Encyclopedia.\ http://en.wikipedia.org/w/index.php?title=Crimea\&oldid=1143508397.$ [Online; accessed 10-March-2023].

Wikipedia. 2023b. Transnistria — Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/w/index.php?title=Transnistria&oldid=1143 [Online; accessed 10-March-2023].

Wikipedia. 2023c. West Bank — Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/w/index.php?title=West%20Bank&oldid=1142 [Online; accessed 10-March-2023].