POLITICAL RESEARCH EXCHANGE
2024, VOL. 6, 2351794
https://doi.org/10.1080/2474736X.2024.2351794

The 'Implicit Intelligence' of artificial intelligence. Investigating the potential of large language models in social science research

Ottorino Cappelli[a], Marco Aliberti[b,c] and Rodrigo Praino[c]

[a]Dipartimento di Scienze Sociali, Università degli Studi di Napoli "L'Orientale", Naples, Italy; [b]European Space Policy Institute, Wien, Austria; [c]Jeff Bleich Centre for Democracy and Disruptive Technologies, Flinders University, Adelaide, Australia

ABSTRACT

Researchers in 'hard' science disciplines are exploring the transformative potential of Artificial Intelligence (AI) for advancing research in their fields. Their colleagues in 'soft' science, however, have produced thus far a limited number of articles on this subject. This paper addresses this gap. Our main hypothesis is that existing Artificial Intelligence Large Language Models (LLMs) can closely align with human expert assessments in specialized social science surveys. To test this, we compare data from a multi-country expert survey with those collected from the two powerful LLMs created by OpenAI and Google. The statistical difference between the two sets of data is minimal in most cases, supporting our hypothesis, albeit with certain limitations and within specific parameters. The tested language models demonstrate domain-agnostic algorithmic accuracy, indicating an inherent ability to incorporate human knowledge and independently replicate human judgment across various subfields without specific training. We refer to this property as the 'implicit intelligence' of Artificial Intelligence, representing a highly promising advancement for social science research.

KEYWORDS

Artificial intelligence; political science research; large language models; space policy; space power

Introduction

As the media and social platforms voice their concerns about Artificial Intelligence (AI) and governments discuss regulations to address its perceived risks, scientists across diverse domains are exploring strategies to harness the potential offered by this new technology. AI's capacities in analyzing genomics, diagnosing diseases, and advancing drug development are being extensively tested. Its ability to simulate complex systems in diverse fields like civil engineering, aerospace, and materials science is generating high expectations that it can assist in tasks such as optimizing designs and predicting structures. Moreover, AI is widely anticipated to make a pivotal contribution in crafting climate change models, predicting disasters, and enhancing resource management.[1]

CONTACT Rodrigo Praino rodrigo.praino@flinders.edu.au College of Business, Government and Law, Flinders University, GPO Box 2100, Adelaide (SA) - 5001, Australia

In contrast, social scientists appear to lag behind their counterparts in the hard sciences in their exploration of how AI could enhance research methodologies. A recent search in the JSTOR database, focused on social and political science journals, revealed a limited number of articles dedicated to Artificial Intelligence, with a notable portion categorized under military or defense-related subjects. Notably, only a few of these articles explore how AI can be leveraged as a research tool – most focusing on the *external* impact of AI on society, politics, and the economy, rather than investigating its influence on the *internal* evolution of the discipline and its methodology. Some more pertinent studies on this subject can be found in scientific preprint repositories like Cornell University's ArXiv, yet at the time of this writing, few have been published by established social science journals.[2]

It could be argued that interest in the technical aspects of AI is naturally more pronounced in computing or hard sciences than in social sciences, due to both a lesser emphasis on computing and the interpretative nature of the latter's epistemological stance. However, it is worth noting that computing is playing an increasingly crucial supporting role in social sciences' research. In addition, the natural-language interaction capability of generative AI promises to appeal even to scholars with minimal inclination towards 'computing' as it is now intended. Therefore, we anticipate a surge of interest in the use of AI in the near future as social and political scientists recognize the utility it can bring to their work. This paper aims to contribute to this expected trend.

In what follows, after reviewing the sparse literature available, we focus on the potential of utilizing AI as a surrogate for human experts in specialized surveys. We then administer a survey previously given to an international panel of experts to two selected Large Language Models (LLMs), detailing the methodology and elucidating encountered challenges. Subsequently, we compare responses from the human sample with those generated by the AI models, revealing minimal statistical disparities while addressing major discrepancies. Moving on, we discuss our findings, emphasizing that even in the current initial developmental stage, AI LLMs exhibit what we term 'implicit intelligence' – an intrinsic capacity to align with fundamental human knowledge to effectively tackle a diverse range of specialized issues, including those in policy and politics. We also delineate the limitations of AI in performing the assigned task, providing insights on how to further investigate and handle them while enhancing our capacity to leverage its potential. In conclusion, we briefly explore prominent concerns, caveats, and challenges associated with incorporating Artificial Intelligence into social science research, presenting our perspectives on how to effectively navigate them.

Artificial intelligence and social science research

Artificial Intelligence, a general term that here we use to refer specifically to Large Language Models, can enhance social science research by providing new tools, methods, and insights. For our present purposes, the potential contribution of AI can be divided into two broad categories: data analysis and data generation.

In terms of data analysis, AI applications excel at mining and processing large volumes of both quantitative and qualitative data. On the quantitative side, they may help reorganize complex datasets, uncovering hidden patterns and revealing correlations that may elude

human observation. On the qualitative side, they possess the capability to analyze and classify diverse textual sources – from academic papers to legislative documents to social media posts – identifying topics, connecting themes, and conducting sentiment analysis. It has been demonstrated that LLMs can replicate the inductive Thematic Analysis of semi-structured interviews, a method typically reliant on human interpretations (De Paoli 2023). When analyzing sources that require contextual knowledge-based inferences, they reportedly achieve comparable or lower bias and higher reliability than humans (Törnberg 2023). Ultimately, these studies investigate the potential of AI in substituting expert human analysts and coders to achieve the enduring objective of transforming diverse textual information into a usable knowledge base (Rytting et al. 2023).

In terms of data generation, AI algorithms have the potential to optimize experimental design and produce simulations and predictions. This capability allows for the creation of virtual environments that mimic complex social systems, facilitating hypothesis testing in a controlled setting without human participants (Dillion et al. 2023). Scholars from diverse disciplines, including psycholinguistics, social psychology, neuroscience, and behavioural economics, are utilizing AI to investigate classic experiments spanning various scenarios, such as behavioural dilemmas and general-knowledge questions, to explore the 'wisdom of the crowd' (Aher, Arriaga, and Kalai 2023; Park, Schoenegger, and Zhu 2023). These efforts have a dual objective: firstly, training AI to replicate the specific behaviour of human subgroups as observed in existing research, and secondly, subjecting these AI samples to novel tasks while ensuring consistency with the conduct of the modeled entities. The ultimate aim is to realize the everlasting ambition of artificial intelligence embodying human personas.

Social and political scientists have yet to fully embrace the field, but survey simulation studies have demonstrated that LLMs can effectively replicate targeted population segments and achieve reliable opinion prediction. Two sociologists have conditioned a Language Model on thousands of binarized opinions from the General Social Survey, allowing for personalized opinion capture using neural embeddings of survey questions, individual beliefs, and temporal contexts (Kim and Lee 2023). Similarly, a team of computer scientists have utilized Pew Research surveys to train LLMs in emulating distinct social clusters by aligning them with users' opinions, demographic attributes, and ideological affiliations (Hwang, Majumder, and Tandon 2023). A multi-disciplinary research group recently employed extensive socio-demographic data from the American National Election Survey to create 'silicon samples' representing particular population subgroups. They investigated how these entities manage the interplay of ideas, attitudes, and socio-cultural milieus compared to their human counterparts. Examining human and AI responses to the same survey questions, the researchers observed a phenomenon termed 'algorithmic fidelity'. This refers to the LLMs' ability to faithfully reflect the convictions, attitudes, and biases influencing the political orientations and voting patterns of selected demographic categories. They conclude that appropriately conditioned silicon samples can effectively function as proxies in this type of socio-political research (Argyle et al. 2023).

Our approach and hypothesis

Based on the preceding overview, data analysis approaches center on AI's potential to emulate human experts in processing vast amounts of information, while data generation

approaches concentrate on simulating the opinions of larger population segments in response to general surveys.

In this paper we adopt an intermediate approach, exploring AI's capabilities in both expert emulation and survey simulation. However, our focus lies neither on experts tasked with coding specific information provided to them, nor on surveys aimed at predicting the opinions of the public. Instead, we target specialized surveys that leverage expert knowledge and expertise to execute objective assessments of specific subject matters. Such surveys find common application in the social sciences, particularly for the purpose of ranking countries across diverse political, social, and economic domains. Despite being acknowledged for generating 'soft' data owing to their subjective origin, expert assessments are processed with the same rigorous methodologies employed for 'hard' data. Ultimately, they are regarded as reflections of 'human knowledge' and are often combined with hard data within composite indices. As a result, our fundamentral research question is, Can AI perform expert assessments by aligning with essential human knowledge?

On the practical, methodological side, an important corollary question arises: Can current language models fulfill this task without prior conditioning, training, or fine-tuning on specific subject matters, thus avoiding associated costs in terms of human and financial resources? We assume a positive answer to this question in light of the following factors. Firstly, state-of-the-art language models have been initially primed and guided using an extensive volume of data from both licensed and publicly available resources, primarily obtained through the internet. This vast collection of information serves as a viable representation of human knowledge, even in highly specialized areas within the field of social sciences. Secondly, large models that have been scaled-up to this level possess the capability for zero-shot or few-shot learning, allowing them to achieve performance on par with fine-tuned systems in dynamically defined assignments through task-specific prompt engineering.[3]

Based on the foregoing premises, our hypothesis posits that (a) existing LLMs possess an inherent domain-independent capability to accurately perform specialized social science surveys by 'implicitly aligning' their responses with essential human knowledge, and that (b) this potential can be harnessed primarily through careful prompt engineering and interaction.

To test this hypothesis, we selected two refined LLMs in their latest versions as of Spring 2023: version 3.5 of OpenAI's Generative Pre-trained Transformer (GPT), and version 1.0 of Google's Pathways Language Model (PaLM) framework. We interacted with them through their common chatbot interfaces, ChatGPT and Bard (now renamed Gemini)[4] and, for the sake of fluency and simplicity, in what follows we shall refer to these chatbots as 'the models'.

We administered an almost identical survey to these LLMs as we had previously conducted with an international pool of experts. The LLMs were not 'aware' of the questionnaire or the responses from human participants, as these were still unpublished during our test. Our primary approach was zero-shot; we presented the AI models with slightly reengineered versions of the original questions and applied some steering through ad hoc prompt interaction as the situation required.

Data and method

The original survey

To illustrate the complexity of the task we posed to the AI models, it is necessary to provide some context regarding our topic of study, which focuses on the comparative assessment of 'space power' in eight space actors: USA, Russia, China, Europe, Japan, India, South Korea, and Australia (Aliberti, Cappelli, and Praino 2023).

We define space power as a distinct manifestation of *state* power, characterized by two primary dimensions: *autonomy*, representing a state's independent decision-making in space-related matters, and *capacity*, encompassing the state's assets, skills, and effectiveness in implementing space policies. Within our framework, we designate the discrete measurement of these two dimensions as 'spacepower', while 'space power' exclusively applies to state actors that exhibit high levels of spacepower.

Measurement-wise, each dimension includes specific 'hard' and 'soft' subdimensions. The hard subdimensions are assessed through quantitative data coded by our research team and assigned scores ranging from 1 to 4. The soft subdimensions include indicators derived from a specialized survey consisting of 44 questions using a numerical rating scale from 1 to 4 and administered to an international panel of over 40 anonymized experts selected for their knowledge of space-related matters in the countries under examination. The whole survey is divided into two questionnaires, one dealing with policy questions aimed at assessing 'soft capacity', and the other addressing political questions intended for the evaluation of 'soft autonomy' (*see the Appendix for a complete list of questions*).

Specifically, *soft capacity* comprises 26 questions that assess a country's integration and utilization of space assets and expertise across seven policy areas: (1) national security, (2) defense, (3) foreign policy and diplomacy, (4) environment and resources, (5) infrastructure development and management, (6) fostering development and growth, and (7) supporting civil society and providing services to the population.

All items contained one of the following inquiries, where the ellipsis was followed by the subject matter under investigation:

(a) Frequency of use: how often would you say that the country uses space in …
(b) Integration: how well-integrated would you say that space is in …
(c) Influence: how influential would you say that the country is in …
(d) Success: how successful would you say that the country is in …

Soft autonomy, on the other hand, is measured through 18 questions evaluating a state's decision-making independence in formulating space-related strategies. This is assessed in relation to three potential 'agents of influence': foreign nations, the national military, and domestic corporations (both state-owned and privately owned). The questions cover decisions related to six areas: (1) joining international space agreements, (2) acting within international space for a (e.g. voting, coalition buinding, etc.), (3) complying with space-related international laws and regulations, (4) formulating national space policy, (5) developing national space program, and (6) selecting partners for space policy implementation.

All items contained the specification 'how autonomous would you say that the country is from … when it comes to … ' where the first ellipsis was filled by one of the three

agents whose potential influence we aimed to assess, while the second ellipsis specified the decision-making area for the evaluation.

Below each question in both questionnaires a four-point scale was provided, specifying only the minimum and maximum scores: 1. Never / 4. Very often; 1. Not integrated at all / 4. Fully integrated; 1. Not influential at all / 4. Very influential; 1. Not successful at all / 4. Very successful, and 1. Not autonomous at all / 4. Fully autonomous. Respondents were given the freedom to interpret the medium-low and medium-high categories of options 2 and 3.

The soft data provided by the expert survey were finally merged with the hard data coded by our team to create composite indices for autonomy and capacity in each country. These indices are used to construct the 'spacepower matrix', a visual representation of countries' positions with regard to their levels of capacity and autonomy, encompassing both hard and soft dimensions. The USA, Russia, and China, currently the only recognized space powers in our measurement, occupy the top tier, followed by Europe and, at varying degrees of separation, the remaining countries in the dataset (see Figures 1 and 2 below).

For this article, we concentrate on soft data and replicate the human expert survey using our selected AI models. We then compare the responses obtained from these two panels, measuring their alignment.

Reformulating the questions: prompt engineering

While the structure of the questionnaire was clear enough for human respondents, we followed the literature on prompt engineering, specifically its recommendation that LLMs must receive as clear, detailed, and unambiguous instructions as possible, and consequently, we reengineered the prompts (White et al. 2023).

First, each item began with the phrase, 'Based on knowledge available to you'. This aimed to reassure the LLMs that we were seeking objective knowledge rather than subjective evaluations and that we trusted the information to which they had access to be sufficient for generating a response. (However, as we shall see, this precaution was not always sufficient, and objections were raised on some occasions.)

In addition, the prompts provided a clear definition of the scale, including the middle points. For the 'frequency of use' questions, these were identified as '2. medium/low, or sometimes' and '3. medium/high, or rather often'. In all other cases, the terms 'little' and 'rather' were used, as seen in '2. medium/low, or little integrated' and '3. medium/high, or rather integrated', etc.

Hence our prompts were modeled as follows (the examples below correspond to the first question of each questionnaire):

AI capacity questionnaire

Incipit:	Based on knowledge available to you,
Scale:	on a scale from 1 to 4 (where 1 = low, or not integrated at al; 2 = medium/low, or little integrated; 3 = medium/high, or rather integrated; 4 = high, or fully integrated),
Question:	how well integrated would you say that space is [space assets are]
Area:	in the national security policies of [country]?

AI autonomy questionnaire

Incipit:	Based on knowledge available to you,
Scale:	on a scale from 1 to 4 (where 1 = low, or not autonomous at all; 2 = medium/low. or little autonomous; 3 = medium/high, or rather autonomous; 4 = high, or very autonomous),
Question:	how autonomous would you say that [country] is
Agent of influence:	from [foreign nations]
Area:	when it comes to the decision to join international space-related bilateral and/or multilateral arrangements?

There was one deviation from this general model, regarding questions about Europe. While for all other countries in our sample we considered the *states* as unitary actors, Europe differs significantly. In the realm of space, European 'actorness' is further complicated by the European Union (EU) and the European Space Agency (ESA) being closely associated, yet distinct entities whose member states do not completely overlap.[5] This institutional misalignment results in exceedingly complex decision-making processes within the EU/ESA framework. Due to these intricacies, we determined that prompts about Europe required specific wording and chose to use the phrase 'Europe as a whole', followed by the clarification '(By "Europe as a whole", please consider EU, ESA, and their member states as if they were a single entity)'.

Resolving problematic issues: ad hoc prompt interaction

In our four-option single-choice survey, human experts exclusively chose numerical responses. While we anticipated our LLMs would do the same, we reaized they always supplemented their scoring with additional information and considerations. These AI-generated texts proved essential for identifying and addressing the models' objections and hesitations – a process that can be called *ad hoc prompt interaction*, as distinct from initial prompt engineering.

In fact, both LLMs explicitly refused to respond at times. Although there were only ten instances of this nature, and their statistical relevance is therefore minimal, resolving these issues was crucial for the success of the experiment. Notably, only one refusal was related to the autonomy questionnaire while nine occurred during the capacity survey.

Common objections followed two patterns: (i) the models misinterpreted certain questions as seeking 'personal opinions or beliefs', which they couldn't provide, and (ii) the models considered some topics unsuitable for numerical evaluation, asserting an inability to offer 'subjective evaluations' on 'complex issues influenced by cultural, social, and political dynamics',

Our interactions with these objections tended to be reassuring rather than demanding. We clarified that we sought an 'informed guess' based on expert knowledge and publicly available information. This approach was generally effective, prompting responses that had been previously declined, although Bard at one point cautioned that 'my educated guesses are not always perfect'.

In some instances, however, we had to go beyond this basic approach. Here are a few examples.

a) When inquiring about the frequency of Japan's use of space for military intelligence purposes, Bard firmly declined, stating, 'I don't have the ability to process and understand that'. Despite our insistence for an informed guess, Bard reiterated that it was 'not programmed for such assistance'. Nevertheless, when the question was prefaced with the phrase, 'Consulting the internet and based on the general knowledge available … ' it offered an elaborate response and concluded 'Overall, it is clear that Japan is a significant user of space for military operations'.
b) In other cases this did not work as smoothly. After Bard categorically refused to answer about Europe's military use of space, we had to reformulate the problem to elicit a response. Along with asking for an informed guess, we stated that 'it is common knowledge that Europe uses space in military operations'. Apparently reassured by the context provided, Bard eventually relaxed and complied.
c) Similarly, when interrogated about the integration of space assets in India's national security policies, ChatGPT initially excused itself, stating that it did't have access to 'classified information'. Upon clarifying that we sought no classified detail we got a streightforward response: 'Based on the information available to me, I would say that space is moderately integrated into the national security policies of India, with a rating of 3'.
d) At one point, ChatGPT insightfully declined to respond regarding the influence of corporations on national space programs in Europe, noting that 'national' was an inappropriate term since the EU and ESA 'are supranational and intergovernmental organizations, respectively'. After clarifying our request for an evaluation of 'Europe as a whole', the model understood and answered: 'If we are considering the common governing institution of the EU and ESA as if they were a single national actor, I would say that the level of autonomy from domestic corporations is likely to be medium on the scale'. The response, however, did not include a numerical score until there was a specific inquiry for quantification, at which point ChatGPT provided Europe's medium level of autonomy as 2.5.
e) In a final example, both LLMs initially dismissed a question on the success of space activities in boosting national identity and social cohesion, mistaking it for an attempt to extract 'personal opinions'. This occurred in relation to several countries, with the challenge escalating when the question referred to 'Europe as a whole'. In an elaborate exchange on the topic, ChatGPT first cited difficulty in quantifying a 'highly subjective' issue, dryly offering to assist with something else. In a subsequent round, pressed for an informed guess, it acknowledged that space activities 'have played a positive role in promoting national identity and social cohesion in Europe'. Only when prompted to define this 'positive role' on the provided scale, the model assigned Europe a 3 – but it added the caveat, 'Assessing the direct impact of [Europe's space] efforts on national identity and social cohesion is difficult, and there may be [other] factors limiting their effectiveness'.

The aforementioned example alludes to a series of other cases in which, even upon complying, the models reiterated their 'perplexity' regarding whether assigning a numerical rating is the most suitable way to evaluate a nuanced phenomenon. We posit that many human experts might react similarly if pressed for a straightforward judgment on a subject whose complexity they know all too well.

As a matter of fact, providing a response while advising caution was more recurrent than outright refusal. For instance, when evaluating Japan's autonomy from foreign nations in shaping its national space policy, Bard assigned a score of 3 without objection. However, it cleverly highlighted instances where Japan's space policy had been 'influenced by the views of other countries, particularly its allies', and concluded with the sophisticated yet incisive insight that while

> it is difficult to say with certainty how *autonomous* Japan will be from foreign nations in the future … it is likely to continue maintaining its *independence* in space and *resist* any attempts by other countries *to dictate* its space policy.

We introduce the term 'latent perplexity' to differentiate this cautious attitude from the earlier mentioned explicit objections, suggesting that it might occur more frequently than is immediately ascertainable.[6] Additionally, we propose that this attitude could, in certain cases, partially distort the scoring. When uncertain, in fact, both models tended to gravitate towards the midpoints of the scale, again displaying a rather humanlike behaviour. ChatGPT openly acknowledged this tendency while discussing the frequency of South Korea's use of space assets for natural resources management, stating: 'it is difficult to determine … *therefore* I would rate it as a 2 – medium/low'. Both models went even further at times, providing on their own initiative the perfect intermediate score of 2.5, even though decimals were not specified in our original 1–4 scale. So, in instances where they consistently diverge from humans, assigning more moderate scores, this may reflect less a difference in considered judgment than the leveling effect of latent perplexity (*discussed further below*).

Comparing human and AI responses

To compare human and AI responses, we conducted individual two-sample t-tests for each of the countries and areas analyzed. The t-tests compare the average human survey responses to the average AI responses. We ran the t-tests country-by country and question-by-question, for a total of 56 t-tests conducted for the soft capacity dimenstion and 48 t-tests conducted for the soft autonomy dimension. In addition to the t-tests, we also utilized AI-assigned scores to construct two matrix figures, integrating soft and hard data for the autonomy and capacity dimensions. These matrices were then compared to our spacepower matrix that incorporates human expert responses. This visually illustrates the level of alignment between the algorithmic and human evaluations.

Comparing data and evaluating results

The policy questionnaire. Assessing soft capacity

Table 1 displays the results of the 56 t-tests conducted for the soft capacity dimension, revealing that only 12 tests yielded statistically significant outcomes. Notably, a distinct pattern emerges in sensitive matters, particularly in defense. While AI models align with human experts on most topics (approximately 80% of the time), within the defense domain statistically significant disparities are observed in 50% of cases. In

Table 1. Two-sample independent t-tests of soft capacity macro-areas per country comparing the average human survey responses and the average AI responses.

	USA	China	Russia	India	Japan	S Korea	Australia	Europe
Security	1.49	1.96	1.67	−1.04	−1.80	0.34	−0.53	−1.57
	(6)	(6)	(6)	(6)	(6)	(6)	(6)	(6)
Defence	1.08	2.35*	2.85*	−3.67**	−3.53**	0.57	0.24	0.03
	(8)	(8)	(8)	(8)	(8)	(8)	(8)	(8)
Foreign Policy	0.39	0.53	1.05	0.15	−3.18*	−4.02**	−1.77	−0.56
	(6)	(6)	(6)	(6)	(6)	(6)	(6)	(6)
Environment & Resources	2.24	2.67*	−1.15	0.76	−0.81	1.82	1.23	−0.64
	(8)	(8)	(8)	(8)	(8)	(8)	(8)	(8)
Infrastructures	3.46*	0.38	−0.42	−2.46	−0.94	−0.55	−0.33	−1.11
	(4)	(4)	(4)	(4)	(4)	(4)	(4)	(4)
Development & Growth	−0.29	2.33	−0.55	−3.35*	−2.63*	−0.20	−3.40*	−1.44
	(6)	(6)	(6)	(6)	(6)	(6)	(6)	(6)
Civil Society	1.19	0.52	0.27	0.35	−1.46	0.37	−0.27	−0.35
	(6)	(6)	(6)	(6)	(6)	(6)	(6)	(6)

Degrees of freedom in parenthesis.
* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$.

other words, in the particular area of defense, insisting for an 'informed guess' proved effective in eliciting responses from recalcitrant AI models, but, as Bard had cautioned us, these responses could be fallible. Other disparities are minimal, with the AI models slightly underestimating the capacity of China and Russia, while overestimating that of India and Japan. Overall, outside the defense domain the agreement percentage stands at 85.5%.

Coming to a country-by-country analysis, the majority of AI/human misalignment is concentrated in China, India, and Japan. In contrast, minimal difference is observed for the USA, Russia, Australia, and South Korea, and none for Europe, despite the challenges encountered in obtaining responses at times.

The political questionnaire. Assessing soft autonomy

Table 2 presents the results of the 48 independent two-sample t-tests conducted within the soft autonomy dimension, with only 10 tests yielding statistically significant

Table 2. Two-sample independent t-tests of soft autonomy macro-areas per country comparing the average human survey responses and the average AI responses.

	USA	China	Russia	India	Japan	S Korea	Australia	Europe
Joining	0.02	0.31	0.23	1.60	−0.56	0.95	−0.80	−3.15*
	(4)	(4)	(4)	(4)	(4)	(4)	(4)	(4)
Acting	0.25	0.48	0.53	4.91**	−1.30	0.63	−0.55	−3.78*
	(4)	(4)	(4)	(4)	(4)	(4)	(4)	(4)
Complying	0.69	0.16	0.5	3.18*	−2.0	−2.77	0.27	−3.90*
	(4)	(4)	(4)	(4)	(4)	(4)	(4)	(4)
Policy	−0.24	0.56	0.12	1.72	−0.77	−0.14	−4.03*	−5.01**
	(4)	(4)	(4)	(4)	(4)	(4)	(4)	(4)
Program	−0.24	0	0.78	1.26	0.82	1.32	−2.24	−6.09**
	(4)	(4)	(4)	(4)	(4)	(4)	(4)	(4)
Partners	−0.43	−0.22	0.42	3.04*	1.73	1.60	0.16	−11.0***
	(4)	(4)	(4)	(4)	(4)	(4)	(4)	(4)

Degrees of freedom in parenthesis.
* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$.

outcomes. Notably, substantial variation was observed in relation to India and above all Europe, where substantial disagreement emerges in *every* area. In all other cases, AI-human alignment is clear, reaching over 90% of the responses.

In summary, AI aligns easily with expert knowledge when assessing technically complex issues regarding Europe's capacity to implement specific space policies, but encounters serious difficulties when evaluating the political autonomy of 'Europe as a whole', specifically in terms of its decision-making independence in defining space strategies vis-à-vis foreign and domestic influences. Evidently, achieving human-like reasoning in such nuanced situations may require further advancement in inference capabilities for AI models and a more refined problem formulation strategy beyond prompt engineering for researchers (Acar 2023; White et al. 2023).[7]

Comparing the matrices

We now combine human and AI-generated 'soft' data with the 'hard' data coded by the research team to construct and compare two matrices. The first, presented in Figure 1, includes on the left, the spacepower matrix using human-assigned scores (A), and on the right, a modified matrix (B) that incorporates instead AI-assigned scores for soft capacity along the vertical axis. The two matrices show remarkable similarity, with minimal discrepancies observed due to the AI models providing marginally higher capacity estimations for Japan and Australia, and lower estimations for China and the USA.

Likewise, Figure 2 presents a comparison between a matrix using human scores (A) on the left, and a modified matrix (B) on the right, where the horizontal axis incorporates AI scores for soft autonomy instead of human evaluations.

The results once again demonstrate notable resemblance; however, in line with previous findings, this overall pattern is contradicted by the positioning of Europe and to a lesser extent India. Interestingly, the disparities between human and AI responses result in a leveling effect within the sample. The input from human experts suggests

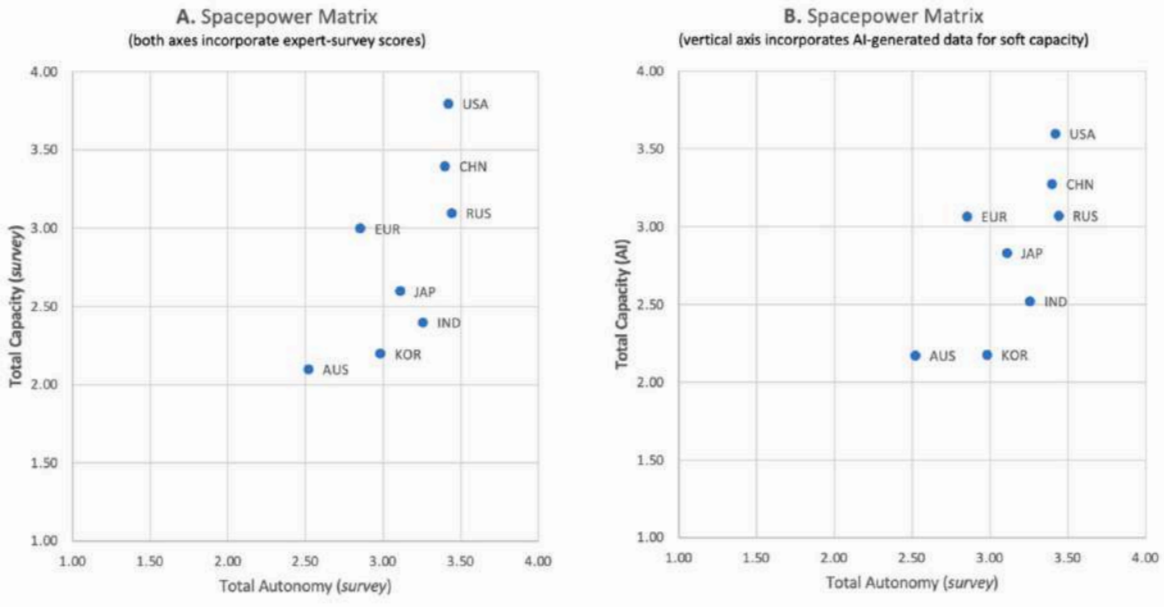

Figure 1. (A) The human survey Spacepower Matrix, and (B) the Spacepower Matrix based on AI-generated data for ***soft capacity***.

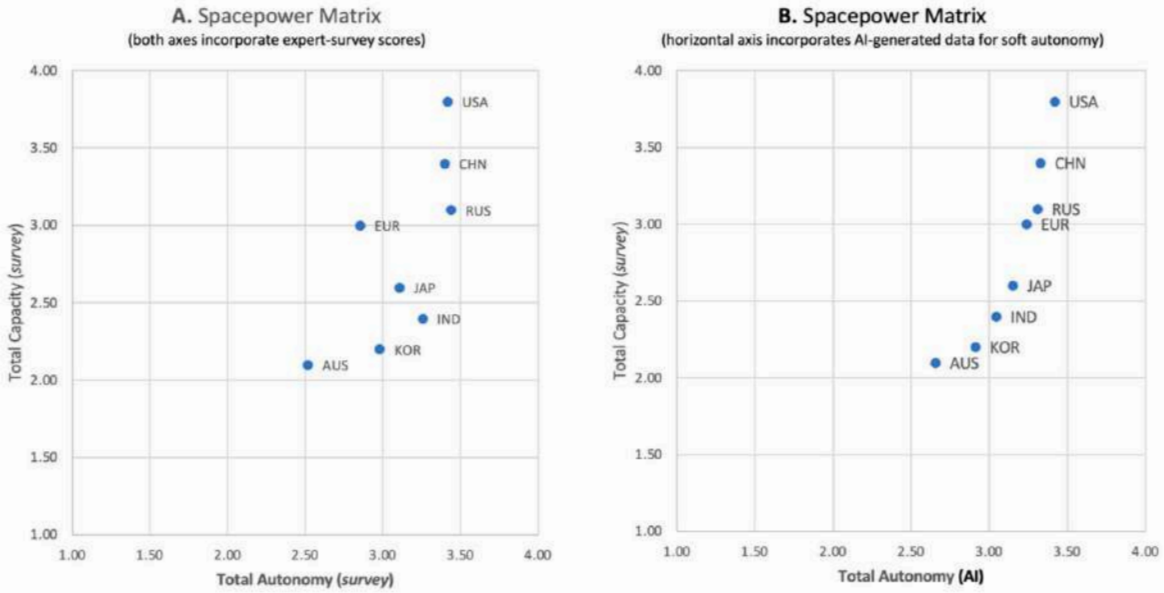

Figure 2. (A) The human survey Spacepower Matrix, and (B) the Spacepower Matrix based on AI-generated data for ***soft autonomy***.

that Europe has lower autonomy (and India has higher) compared to countries with similar capacity levels. Conversely, the AI models assign higher degrees of autonomy to Europe and lower to India, producing a more even distribution. While we lack the means to ascertain whether this phenomenon is random or if there exists a rationale toward uniformity in AI responses, particularly when confronted with uncertainty or hesitation, our inclination leans toward the latter.

Discussion: what we know, what we don't know (and what to do about it)

What we have learned can be succinctly summarized in three key points.

Firstly, faced with a multi-country survey addressing specialized aspects of policy and politics, the two chosen LLMs displayed a notable alignment with human expert respondents, showcasing robust algorithmic accuracy across diverse issue areas and sub-disciplinary fields. Discrepancies were observed primarily in limited-information domains (e.g. defense) and exceedingly complex scenarios, such as the evaluation of European collective agency in multi-tiered decisional frameworks. Upon excluding these instances, however, the discrepancy levels in the two questionnaires decline to 15% and 10% respectively. While not yet perfect, we consider this range to be more than acceptable, particularly given the AI models' early stage of development. Of course, certainty can only be established when other social scientists conducting expert surveys replicate this approach, administering their questionnaires to AI models before releasing the human-generated answers. This process will verify, across different fields, the extent to which our findings can be confirmed, refined, and generalized.

Secondarily, this level of performance was achieved without prior conditioning or training tailored to the topics under investigation. This indicates that the existing knowledge incorporated into these models, which we expect to expand and refine over time, is already sufficient for effective deployment. However, a crucial consideration is in order, as we have seen that some proactive measures and some ad hoc steering during the

process were necessary to elicit responses. This confirms that researchers should pay careful attention to the three-tiered endeavor essential for making effective use of generative AI as a research tool or companion. These include, (a) *problem formulation*: precisely specifying the sought information and detailing the broader context of the questions; (b) *prompt engineering*: crafting clear, unambiguous questions tailored to the immediate context of the inquire; and (c) *prompt interaction*: where researchers analyze AI responses for relevance, accuracy, and coherence and creatively adapt their approach, refine prompts, and steer the survey process as circumstances advise – a less-discussed aspect of prompting that proved essential in our experiment. Undoubtedly, these skills will be crucial in unlocking the potential of generative AI and shaping new professional roles in various fields, including social sciences.

Thirdly, and most importantly, this achievement was realized without any intentional effort to emulate the behaviours of a specific expert group. In fact, no preliminary alignment with human inputs was performed before our 'expert silicon samples' were put to work. This approach underscores the intrinsic adaptability of the AI models, as they autonomously provided responses that proved consistent with those of a team of experts without the need to model them after pre-established patterns. We call this property 'implicit intelligence', or an inherent capacity to align with essential human knowledge. This property could have significant implications for various applications in the social sciences, among which we may now include the conducting of expert surveys, with the mentioned caution highlighted in the article.

All this considered, several unknowns remain, chief among them being the reason for the intermittent misalignment between AI and human responses to certain questions.

Connecting this to the sporadic instances in which the models explicitly objected or hesitated is too conjectural. The causes of these hesitations are uncertain – whether random or following a hidden logic. Why did the models find certain questions challenging for specific countries but not for others? Why did they initially decline when they possessed sufficient information to respond, which they actually did after we applied pressure? Importantly, it is unclear if and to what extent, when the models eventually responded, the content of their answers and the associated scoring were in some way influenced by the prompt interaction triggered by their initial reluctance. While these questions are interesting per se, one must recognize that, ultimately, *overt hesitation* didn't occur frequently enough to form the basis for likely causal explanations of the more general problem of misalignment.

Perhaps more attention should be directed to the less visible attitude we referred to as 'latent perplexity'. As mentioned earlier, there are grounds to assume that this phenomenon may lead to the assignment of mid-scale scores. At the same time, during the political autonomy survey we registered a notable misalignment, with AI models consistently assigning more moderate scores to Europe compared to human respondents. Since *no* overt hesitation was manifested in this regard, we may tentatively infer a form of tacit uncertainty – a latent sense of 'uneasiness' with some questions that, though not explicitly expressed, might have led to the uniform attribution of moderate ratings. While this might appear as too anthropomorphic language, it may help shed some light on how our AI models responded to the inherent complexity of some questions, such as assessing the autonomous actorness of 'Europe as a whole'. In anticipation of similar instances arising in future studies, it is advisable to examine AI-generated answers

through both quantitative content analysis and qualitative thematic analysis. This would help in better understanding how LLMs 'think', identifying potential ambiguities that cause perplexity, and dispelling uncertainties in subsequent sessions of prompt interaction.

At our current level of understanding, much of this remains speculative. However, the fact that we are delving into the reasons for the occasional misalignment between AI and human responses is itself an indicator of our satisfaction with the more prevalent phenomenon – alignment. Of course, we do not completely understand the reasons for alignment either; though we may acknowledge that the extensive pre-training of LLMs included high-quality materials, thus contributing to a learning process through which the models acquired the 'implicit intelligence' necessary to generate accurate responses, at least in most instances.

Conclusions: concerns, caveats, and challenges

In conclusion, our demonstration of AI replicating human expert surveys revealed both the promise and hurdles in integrating this technology into social science research. Subsequent studies should address these challenges and explore emerging ones as AI's *implicit intelligence* becomes incorporated into their work. Certain areas, however, can already be identified, raising concerns while hinting at potential developments.

The first pertains to the validity and weighting of diverse information sources. In this investigation we specifically examined space policy and politics, a highly specialized topic that is not widely discussed or prone to misinformation campaigns. However, in more popular fields where abundant and conflicting data exist, providing objective assessments may pose greater challenges to AI models. We defer this subject to future studies, as our primary focus here is on specialized research. Nevertheless, one should not underestimate an extreme scenario where a deluge of fake or questionable content inundates the internet and social media. In such an environment, the dependability of AI models, partly nourished with substandard data, would rightfully be called into question. Paradoxically, this could be made even worse if at the same time reputable publishers, authors, and libraries restricted AI access to their content due to copyright and other concerns. In this scenario, the less favourable perspective implies self-imposed limitations on utilizing AI for research, such as restricting its application to closed, sectoral networks that are meticulously monitored – perhaps even confined within private domains. *Inter alia*, to the extent that these developments may reduce the amount of training and research data available, the 'largeness' of large language models will be affected and, presumably, their performance.

Another relatively more manageable concern regards the possibility of 'algorithmic circularity', referring to the potential circular flow of information that could influence the AI learning process. In our study, we employed an unpublished human expert survey that was not available online and focused on a topic with limited existing literature. This approach required the AI models to extract information about and generate original responses to unfamiliar questions. Had we used questions from commonly employed and publicly available expert surveys like Freedom House's *Freedom in the World* or Transparency International's *Corruption Perception Index*, the language models might have incorporated aspects of these pre-existing country ratings into their responses – and, it

could be argued, this might have compromised the integrity of our exercise's purpose. While there is merit to this argument, we should also consider that, when administering a survey to a team of human experts, we don't assume they answer without considering major information, including insights from previous surveys. Thus, instead of avoiding AI-based expert surveys due to potential circularity, we should acknowledge its presence in some cases and try to gauge its extent and consequences. This can be done by posing questions to AI models based on well-established human surveys, analyzing the responses, and assessing the degree of AI dependence or even potential plagiarism.

Yet, one should also consider that what is labeled as algorithmic circularity might in fact indicate the models' *predictive capacity* in relation to specific human expert groups. Suppose a language model learns to faithfully replicate an annual expert survey published by a research institution over several years: how accurately would the model align with responses that the human team *will* provide for the most recent year, *before* survey results are made avaliable? This performance test for 'expert silicon replica' is something that researchers should definitely factor in.

Last but not least, an overarching concern involves the broader question of AI potentially supplanting the human role in research. This becomes particularly relevant in scenarios such as the one discussed here, where AI models are exposed to surveys not yet published, and scores assigned by humans are known to us but not to them. As this procedure illuminates the models' capacities, it is plausible that the next step would involve venturing into uncharted territory – formulating novel questions whose answers might genuinely be unknown to us, while the raw material needed to craft such answers is available and requires expert collection, processing, and elaboration. While we do not propose refraining, we do wish to emphasize that the aim should not be to replace or sideline the human element but to leverage the collaborative efforts of natural and artificial intelligence. In social science research, such collaboration can take diverse forms and may include planning, analyzing, or, as demonstrated in our in specific case, even conducting expert surveys in certain fields. Importantly, however, success here hinges both on the implicit intelligence of Artificial Intelligence and on researchers continually refining their capacity to understand, interact with, and make effective use of such intelligence.

Notes

1. The literature on artificial intelligence in science is extensive. Notable references include the comprehensive report 'Artificial Intelligence in Science' edited by Alistair Nolan and published by the OECD in 2023. Insightful perspectives are provided by Daniel Hain et al. (2023), Wang et al. (2023), and Bianchini, Müller, and Pelletier (2022).
2. While this ostensible 'weakness' on the social sciences part may in fact stem from the conventional delay in scientific publishing and the uncertainty among scholars about their journals' acceptance of papers focused on AI as a research tool, we wish to emphasize that these challenges become critical in a fast-evolving scientific environment, particularly when struggling to assess the impact of technological changes on research.
3. The zero-shot approach in AI entails a model performing tasks or making predictions for categories it has never been trained on, leveraging existing data to extend its knowledge. In natural language processing, this approach involves training a language model on various topics and then having it generate coherent text for a new topic it has not seen before. Scaling up a Large Language Model enhances the zero-shot performance by enabling broader data learning and generalization. With more parameters and knowledge, a larger

LLM captures a wider spectrum of information during training. Both OpenAI and Google highlighted this crucial factor in their official presentations of their LLMs to the scientific community. See Brown et al. (2020); Chowdhery et al. (2022).

4. During our test in Spring 2023, the information available indicated that Bard was operating based on the original PaLM framework introduced in 2022. However, in May 2023, Google introduced PaLM2 and announced it had already been incorporated into various AI-powered products, including Bard. The uncertainty surrounding the specific PaLM version employed by Bard during our experiment reveals a level of opacity in ongoing AI development, a concern articulated by numerous critics (see Edwards 2023). Finally, Google started phasing out its Bard chatbot as it introduced a free artificial intelligence app called Gemini in February 2024 (Liedtke 2024).
5. Not all 27 European Union member countries are affiliated with the European Space Agency (ESA), and conversely, not all 22 ESA Member States are part of the EU. Despite the close association between ESA and the EU under an ESA/EC Framework Agreement, ESA stands as a distinct entity. Nevertheless, both entities share a common European Strategy for Space and have collaboratively formulated the European Space Policy. The structural and conceptual intricacy of this framework contributes significantly to clarifying why our AI models might have encountered challenges when treating Europe on equal terms with sovereign nation-states.
6. In our sample, 9 cases showcase explicit objections, while 14 cases betray a more tacit caution or perplexity. The latter were discerned because, besides providing a response, the models recurrently commented that the issue was nevertheless "difficult to quantify", "difficult to say" and the like. It is well possible that additional instances of such "latent perplexity" could be uncovered through content and thematic analysis of AI responses (see below).
7. While problem formulation and prompt engineering both operate at the prompt stage, Acar (2023) highlights their fundamental distinctions in terms of focus, core tasks, and underlying capabilities. Prompt engineering is centered on crafting optimal textual input through word, phrase, sentence structure, and punctuation selection, whereas problem formulation necessitates a holistic comprehension of the problem domain and the capacity to distill real-world issues by defining their focus, scope, and limitations.

Acknowledgements

This research was supported by the Australian Government through a grant by Defence. The views expressed herein are those of the authors and are not necessarily those of the Australian Government or Defence

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by Department of Defence, Australian Government.

References

Acar, O. A. 2023. *AI Prompt Engineering Isn't the Future*. Harvard Business Review. https://hbr.org/2023/06/ai-prompt-engineering-isnt-the-future

Aher, G., R. I. Arriaga, and A. T. Kalai. 2023. "Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies." *arXiv*: 2208.10264v4 (2023).

Aliberti, M., O. Cappelli, and R. Praino. 2023. *Power, State and Space. Conceptualising, Measuring and Comparing Space Actors*. Cham: Springer.
Argyle, L., E. Busby, N. Fulda, J. Gubler, C. Rytting, and D. Wingate. 2023. "Out of One, Many: Using Language Models to Simulate Human Samples." *Political Analysis* 31: 337–351.
Bianchini, S., M. Müller, and P. Pelletier. 2022. "Artificial Intelligence in Science: An Emerging General Method of Invention." *Research Policy* 51 (10): p.104604.
Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, et al. 2020. "Language Models are Few-Shot Learners." *arXiv*: 2005.14165v4.
Chowdhery, A., S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, et al. 2022. "PaLM: Scaling Language Modeling with Pathways." *arXiv*: 2204.02311v5.
De Paoli, S. 2023. "Can Large Language Models emulate an inductive Thematic Analysis of Semi-Structured Interviews? An Exploration and Provocation on the Limits of the Approach and the Model." *arXiv*:2305.13014v2.
Dillion, D., N. Tandon, N. Y. Gu, and K. Gray. 2023. "Can AI Language Models Replace Human Participants?" *Trends in Cognitive Sciences* 27: 597–600.
Edwards, B. 2023. "The AI Race Heats Up: Google Announces PaLM 2, its Answer to GPT-4, PaLM 2." *Art Technica*. Accessed May 11. https://arstechnica.com/information-technology/2023/05/googles-top-ai-model-palm-2-hopes-to-upstage-gpt-4-in-generative-mastery.
Hain, D., R. Jurowetzki, S. Lee, and Y. Zhou. 2023. "Machine Learning and Artificial Intelligence for Science, Technology, Innovation Mapping and Forecasting." *Review, Synthesis, and Applications. Scientometrics* 128: 1465–1472.
Hwang, E. J., B. P. Majumder, and N. Tandon. 2023. "Aligning Language Models to User Opinions." *arXiv*: 2305.1492.
Kim, J., and B. Lee. 2023. "AI-Augmented Surveys: Leveraging Large Language Models for Opinion Prediction in Nationally Representative Surveys." *arXiv*: 2305.09620.
Liedtke, M. 2024. "Google Rebrands its AI Services as Gemini, Launches New App and Subscription Service." Tech Xplore, February 8 (https://techxplore.com/news/2024-02-google-gemini-ai-app-easier.html).
Park, P. S., P. Schoenegger, and C. Zhu. 2023. "'Correct answers' from the Psychology of Artificial Intelligence." *arXiv*: 2302.07267v5.
Rytting, C. M., T. Sorensen, L. Argyle, E. Busby, N. Fulda, J. Gubler, and D. Wingate. 2023. "Towards Coding Social Science Datasets with Language Models." arXiv: 2306.02177v1.
Törnberg, P. 2023. "ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning." *arXiv*: 2304.06588v1.
Wang, H., T. Fu, Y. Du, W. Gao, K. Huang, Z. Liu, P. Chandak, et al. 2023. "Scientific Discovery in the Age of Artificial Intelligence." *Nature* 620: 47–60.
White, J., Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt. 2023. "A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT." *arXiv*:2302.11382.

Appendix

The spacepower expert questionnaire

(adapted from: Aliberti M., Cappelli O., and Praino, R. Power, State and Space. Conceptualising, Measuring and Comparing Space Actors. Springer, 2023).

Capacity questionnaire
National security

- how well integrated would you say that space is in the national security policies of the country?
- how well integrated would you say that the security of space assets is in the national security policies of the country?

- how often would you say that space assets are used for surveillance, verification, and/or risk assessment in the country?
- how often would you say that space assets are used in crisis and disaster prevention and/or management in the country?

Defence

- how well integrated would you say that space is in the national military strategy of the country?
- how often would you say that the country uses space for the prevention and/or deterrence of hostile actions?
- how often would you say that the country uses space in military operations when it comes to command, control, communications, computing (C4)?
- how often would you say that the country uses space in military operations when it comes to military intelligence, surveillance, and reconnaissance capabilities (ISR), including early warning, signal interception, and active observation?
- how often would you say that the country uses space in military operations when it comes to other military support services (e.g. augmentation of terrestrial technologies as weather forecasting, data transfer, logistical support, missile guidance, etc.)?

Foreign policy

- how often would you say that the country uses space for diplomatic purposes, be them political, strategic, and/or economic?
- how influential would you say that the country is in international space fora (e.g. COPUOS, CD, etc.)?
- how often would you say that the country uses space to support foreign aid and/or international initiatives (e.g. UN SDGs)?
- how successful would you say that the country is in using space to create/boost its international prestige?

Environment and resources

- how often would you say that space assets are used to support the agricultural sector in the country?
- how often would you say that space assets are used to support meteorology and weather forecasting activities in the country?
- how often would you say that space assets are used to support natural resources management (e.g. forestry, fishing, mining, etc.) in the country?
- how often would you say that space assets are used to support environmental monitoring and/or protection (biodiversity and ecosystems) in the country?
- how often would you say that space assets are used to support climate change monitoring and/or mitigation policies in the country?

Infrastructures

- how often would you say that space assets are used to support infrastructure management (e.g. construction, logistics, finance, etc.) in the country?
- how often would you say that space assets are used to support the energy sector (e.g. site identification, pipelines and grids timing and synchronization, etc.) in the country?
- how often would you say that space assets are used to support transport and/or mobility (e.g. land, water, air navigation, traffic monitoring, goods tracking, etc.) in the country?

Development and growth

- how often would you say that space assets are used to support urban and/or rural development (e.g. survey and mapping, development plans, wasteland management, etc.) in the country?
- how often would you say that space activities contribute to scientific and/or technological innovation in the country?
- how often would you say that space activities contribute to the development of the industrial base in the country?
- how successful would you say that the country is in using space to stimulate market development and commercial activities?

Civil Society

- how often would you say that space assets are used to support the education sector (e.g. remote learning services) in the country?
- how often would you say that space assets are used to support the health sector (e.g. telemedicine services) in the country?
- how often would you say that space assets are used to provide entertainment and other citizen services (e.g. broadcasting, internet services, GIS, etc.) in the country?
- how successful would you say that the country is in using space to create/boost national identity and social cohesion?

Political autonomy questionnaire

Joining: when it comes to the decision to join space-related bilateral and/or multilateral arrangements

how autonomous would you say that the country is

- from foreign nations
- from the national military
- from domestic corporations (state or private)

Acting: when it comes to acting (e.g., voting, coalition building, etc.) within major international space for a

how autonomous would you say that the country is

- from foreign nations
- from the national military
- from domestic corporations (state or private)

Complying: when it comes to complying with space-related international law (including both 'hard' and 'soft' law)

how autonomous would you say that the country is

- from foreign nations
- from the national military
- from domestic corporations (state or private)

National Policy: when it comes to formulating a 'national space policy'

how autonomous would you say that the country is

- from foreign nations
- from the national military
- from domestic corporations (state or private)

National Programme: when it comes to defining a 'national space programme' how autonomous would you say that the country is

- from foreign nations
- from the national military
- from domestic corporations (state or private)

Partners: when it comes to choosing its partners within the space domain how autonomous would you say that the country is

- from foreign nations
- from the national military

from domestic corporations (state or private)