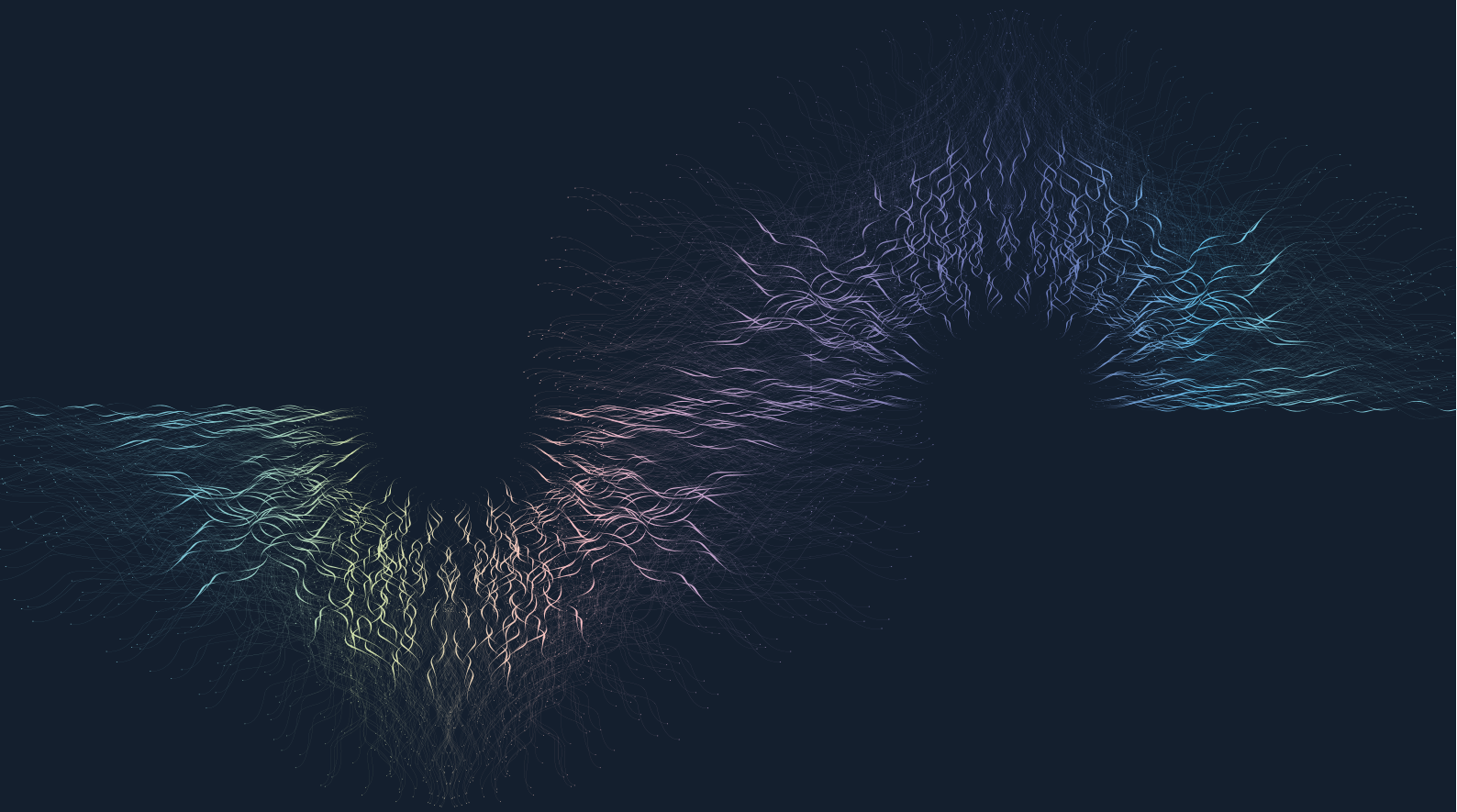
intelcomp

www.intelcomp.eu



Policy brief on the use of Artificial Intelligence (AI)-based tools for Science, Technology and Innovation (STI) policy

AI-based tools for STI policy: asking the 'right' questions for any domain

Why invest in AI tools for STI policy?

The European Commission (EC) vision is to create a new European Research Area (ERA) responding to the future needs, demanding greater (inter) connectivity, digitalization and improved evidence-based and data-driven intelligence gathering to inform policy making and benefit society as a whole. However, for STI policy makers it is a continuous challenge to: 1) select the type of evidence (in the form of both raw data and indicators) needed for any policy decision, 2) identify suitable sources, where it can be drawn from and 3) effectively use methods and tools to exploit the data and generate indicators. It is believed that open data and AI-based tools allow for better, more accurate and more efficient policymaking.

What are the reasons justifying this belief? Looking at the last two decades we observe that the collection and processing of STI data has been facilitated and proliferated. A large share of it is now available online in the form of open data suitable sources (e.g. OpenAIRE)

As AI-based tools will facilitate the third step substantially, the IntelComp approach is to demonstrate the value of using open data sources and new digital tools to generate new indicators and prepare the ground for human decisions.

What for? What are the questions that STI policy makers do not have answers to and why getting more than the traditional information matters?

STI policy questions are hard to answer partly because of the systemic nature of innovation and, therefore, the need to identify proxies of multiple functions of the innovation system which are only partly quantifiable, and analyse their interactions. Time also significantly complicates the work of STI policy makers given the rapid technological evolution on the one hand and the time lag in the production of scientific research outputs and outcomes, which implies the need for both regular and longer term monitoring. Multi-actor involvement in STI (including industrial actors, educational actors, consuming actors, intermediary actors, political and policy/regulatory actors and infrastructural actors) equally challenges data collection but also the understanding of modalities with which they shape innovation processes.

This is best illustrated by an example. In the table below we focus on the key question "Where should resources be invested (individual companies, sectors, value chains) to support the national innovation system to successfully undertake STI and compete internationally?". This is a question of relevance for policy makers during the agenda setting phase of the policy cycle and, while it centers on entrepreneurship, there are relevant aspects to be considered related to knowledge creation, resource mobilization and societal legitimacy. What the example showcases is the complexity in the technical requirements, considering the variety of taxonomies (e.g., sectors, technologies, codes for patents, subject areas for scientific publications, themes for projects) and concordances to link them to one another, which are not readily available. It equally demonstrates the large number of data sources that can be jointly considered which ranges from traditional to alternative less explored sources of data.

Policy cycle phase: Agenda Setting (i.e. Intelligence gathering, problem identification)

Key question: Where should resources be invested (individual companies, sectors, value chains) to support the national innovation system to successfully undertake RDI and compete internationally?

Rationale A: Understand which companies are active in emerging fields (emerging field defined under Knowledge Creation) and likely to excel in the future, this is where you want to invest

Rationale B: Understand where local companies have an RDI specialisation (the answer to these questions will be prepared during the monitoring and evaluation part of the cycle)

Innovation System Function - Entrepreneurial activity

Related Policy questions

- Are companies adapting to technological transformation trends in their respective sectors? How do they compare with major (international) competitors?
- Which companies emerge with specific disruptive technologies in the country/ macroregion/ region/city?
- Are companies emerging with specific disruptive technologies scaling up?
- Are scale ups leaving the country/ macroregion/ region/ city?
- Does the country/ macroregion/ region/ city attract entrepreneurial talent?
- Who are the persistent innovators in the country/ macroregion/ region/ city?
- In which R&D fields do the persistent innovators invest?
- In which R&D fields is the highest share of all company R&D investments?
- In which R&D fields is the country improving its revealed comparative advantage?

Concepts (provisional)
Company RDI activity
Company technology uptake
Company RDI investment
Company RDI funding received

Sources (provisional)
Traditional sources
EU Industrial R&D Investment Scoreboard
Community Innovation Survey (CIS)
Patstat
FP1 – FP7, Horizon 2020, Horizon Europe

Other sources
Company websites
Crunchbase[1] / Dealroom[2] / Eutopia[3] / Cleantech[4] / Pitchbook[5]
Incubator, Accelerator, Investment angels websites
Companies issuing Initial Public Offering (IPOs)
Mergers & Acquisitions (Zephyr, Crunchbase, etc.)

Technical Requirements (provisional)
Temporal evolution
Geographical levels
Taxonomies of sectors, technologies, scientific fields, IPC codes

Associated innovation system functions

Related Policy questions

- Innovation System Function - Knowledge creation: Which scientific fields demonstrate the highest growth in terms of publications/ citations globally? Distinction to be made between basic and applied research (distinction between interdisciplinary publications, basic research and applied research) (using journal classification?/calls for proposals); Which are the emerging interdisciplinary fields globally?
- Innovation System Function - Resources mobilization: What is the size of resources needed to become competitive in each emerging technology?
- Innovation System Function - Creation of legitimacy/ counteract resistance to change: Is resistance expected? Where? Why? How?

1 See: https://www.crunchbase.com
2 See: https://dealroom.co
3 See: https://www.eutopiagreen.com
4 See: https://www.cleantech.com
5 See https://pitchbook.com

Using AI based tools on text available in some of these data sources (corpus) allows dynamic tracking of large amounts of heterogeneous data over different time frames. It facilitates comprehensive overviews of the STI landscape allowing geographic, thematic, sectoral, technological and time evolution analyses. It also allows inter-corpus comparison between sets of documents from different corpora as there is no need for predefined taxonomies. Retrieval and matching with similar publications, patents or projects is possible bringing them together in an aggregated view. All these functionalities save time, effort and cost and ensure more objective, evidence-based assessment is possible.

Can we do it now? Do we have the technologies? What are the challenges these technologies face?

To facilitate the analysis of text data Natural Language Processing and AI techniques are needed. These techniques include topic modelling, conceptual indexing of documents, word embeddings and other forms of document representation. There has been use and experience in existing tools for STI policy like Data4Impact[6] and Corpus Viewer (OECD, 2018)[7], and in several studies (e.g. OECD, 2021)[8]. The technical solutions that need to be incorporated in different components of the IntelComp Platform are thus feasible.

One challenge is the availability of open data suitable sources. While large volumes of public information are available online (think of digital platforms or websites of incubators/accelerators) the access to it is not always encouraged (platforms discourage web crawling or explicitly prohibit it in their terms of service clauses) and the legal grounds permitting web crawling are not always clear. National funding microdata on beneficiaries and projects are equally not always public information. Moreover, for each database inherent biases need to be addressed before concluding on their suitability for STI metrics.

6 http://www.data4impact.eu/
7 OECD (2018), "The digitalisation of science and innovation policy", in OECD Science, Technology and Innovation Outlook 2018: Adapting to Technological and Societal Disruption, OECD Publishing, Paris, https://doi.org/10.1787/sti_in_outlook-2018-17-en
8 Yamashita, I., et al. (2021), "Measuring the AI content of government-funded R&D projects: A proof of concept for the OECD Fundstat initiative", OECD Science, Technology and Industry Working Papers, No. 2021/09, OECD Publishing, Paris, https://doi.org/10.1787/7b43b038-en

Another challenge is to incorporate experts' knowledge in the AI models. Objective measures that are usually used by Machine Learning practitioners are not always correlated with model interpretability, and designing strategies to better align the models to the needs of final users (e.g., identifying subsets of documents related to a domain of interest, reducing model variability, etc.) is of major importance to make users build confidence and ultimately rely on conclusions derived from the models.

A third challenge is to build the right user visualization tools to make the outputs of the models easy to use and actionable to the final users. For this, co-creation methodologies can be exploited, jointly revising user requirements and improving the developed tools in an iterative manner.

How do we start?

Before answering to these challenges, we need to define STI policy needs tailored to the needs of specific domains (AI, climate change and health). In the case of IntelComp we:

- Start with an extensive domain-agnostic long list of questions that STI policy makers may need; key questions cannot receive simple answers, hence, they were subdivided into more concrete partial questions
- Include in these questions all the components of an innovation system
- Identify the data sources, methods of analysis and indicators, which may be relevant to be able to respond to (many of) these questions, prioritizing them according to the policy makers and stakeholders points of view in the specific domains.

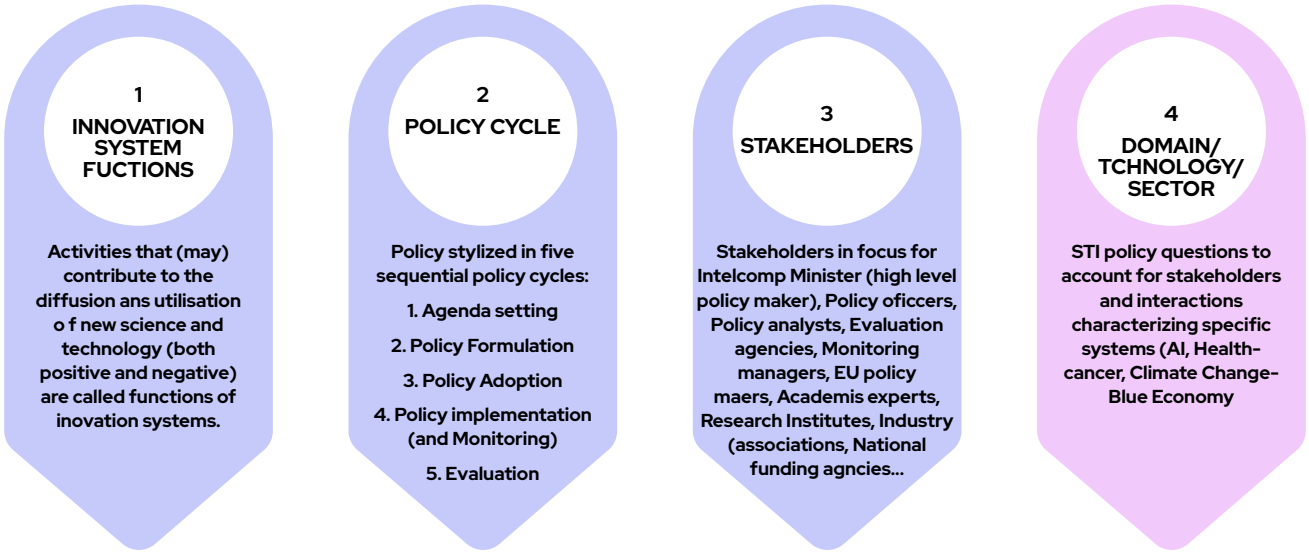
To make the conceptual framework sufficiently practical to be able to frame questions relevant for policy makers, a structure was selected, which is both user-friendly and responds to actual needs. The structure combined two dimensions:

· The type of functions policy makers wish to address within an innovation system

· The moment they are needed within a policy cycle

The components accounted for in the definition of policy questions are summarized below.

Public Administrations as well as many private companies can benefit greatly from AI technologies as well as cloud-deployed services that IntelComp will foster under this action. This potential has not yet been realised, making it a goal of EU STI policy making.



We use all three dimensions... but not all possible combinations to create a basic set of questions

intelcomp

www.intelcomp.eu

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101004870.