ie UNIVERSITY

CENTER FOR THE GOVERNANCE OF CHANGE

JUNE 2023

DATA POLICY: A CONCEPTUAL FRAMEWORK

cgc.ie.edu

CONTENTS

AUTHOR:
Andrea Renda

INTRODUCTION: ON THE ELUSIVE ECONOMICS OF DATA

For over a century, economists have struggled to incorporate information and data in their conceptual frameworks, mostly without succeeding. This is due to the chameleonic features of information and data as economic concepts, which often clash with the otherwise well-established tenets of neoclassical market economics. Not surprisingly, data has been analysed under a myriad of perspectives and defined as a public good, a common good, a club good, a "semicommons", a form of infrastructure, a form of labour, a form of capital, and more. This confusion is understandable, given that, as will be explained below, data can take all these forms, depending on the context and the eye of the beholder. However, the ongoing *querelle* as to the nature, value and specific dynamics of data does not help policymakers when it comes to designing appropriate policies to govern the flow of data in the modern, information-rich world.

So, what do we know about data?

First, we know that it is *special* from an economic perspective. Special in the sense of *species*, in Latin, which denotes "a particular sort, kind, or type". In other words, **data "behaves" neither like ordinary products or services nor like pure public goods. It can be rival as well as non-rival, excludable as well as non-excludable, and it can feature widely different utility functions depending on the type of data considered, as well as the context in which it is used** (see Section 1 below). Data is special also from a legal perspective: it cannot formally be "owned" in the traditional legal sense of the word, yet often circulates based on the decisions of those that enjoy a *jus excludendi omnes alios* over its distribution, and as such it relies on quasi-property rights, often protected technologically more than legally, and sometimes protected through liability rules (i.e. remunerated compulsory access), rather than property rules (i.e., the right to exclude).[1]

Second, data exhibits *quantum* characteristics, in that the same data asset is often "multi-status" (akin to "superposed" in quantum physics) and "multi-purpose" (akin to entangled).

> Just like in quantum physics, the state and position of a given object reveals itself only at the time of observation, and for the specific observer, data can show a different face, value, status and purpose, depending on when and how it is observed, as well as who is observing.

The same data asset can be seen as valueless and invaluably important, depending on the purpose of data collection, and much data acquires value only when aggregated with similar data. At the same time, organising a transaction, such as exchange or sharing, related to data becomes very difficult, since one party may see data as valuable when it is not shared, whereas the other party may be interested in the value of the same *datum* when aggregated with many other data. Hence, the traditional economics of allocative efficiency, based on the idea that comparing the parties' willingness to pay leads to mutually beneficial (Pareto-efficient) contracting, does not necessarily hold in the economics of data sharing and exchange. Moreover, as originally explained by Kenneth Arrow, things are complicated by informational asymmetries and bounded rationality; as well as by the so-called "paradox of information":[2] in a transaction over data, to fully appreciate the value of a data item to acquire or collect, a party should be able to observe it first; but once access to the data item has been granted, the transaction has already taken place. Arrow's information paradox explains why the fear of appropriation can lead to dramatic under-sharing of data, and to a large extent, it laid the foundation for public intervention to reduce transaction costs in business-to-business (B2B) data-sharing (see Section 2 below).

> The value of data critically depends on what the data refers to or is attached to. (...) There is no such thing as data per se; data is always about something.

Third, the value of data is essentially *subordinate*, or in other words, *ancillary* to the value of an underlying asset. This means that the value of data critically depends on what the data refers to or is attached to. In information economics, this has generated a separate stream of literature, which sees the value of information and data as drivers of more efficient transactions and market dynamics rather than economic assets *per se*. Data can relate to the occurrence of events, the recurrence and phenomenology of specific behaviours, the attributes of a product, and much more. There is no such thing as data *per se*; data is always about something: especially in the digital age, very often the collection of data about an asset or behaviour enables either the capture of part of that underlying asset's value or enables the creation of new value through data aggregation. The "datafication" of the economy, powered by the emergence of cyberspace, has institutionalised the decoupling between *value creation,* normally attributable to those entities that produced the underlying asset/event the data refers to; and *value capture*, typically associated with the collection, aggregation and re-use of such data.[3]

Fourth, the value of data is very often dramatically *time-dependent*. While certain types of data may preserve their value over time, some lose part of their value very rapidly, sometimes in fractions of a second. The best example is provided by the millions of dollars spent every year by investment banks to shave microseconds off their high-frequency trading operations. Similarly, data on the current state of a road or rail infrastructure has value when available in real-time. In both cases, data can later be aggregated and re-used to generate predictions and statistics, but the portion of value that was meant for immediate decision-making gets depleted in less than the blink of an eye.

> Thus, in order to estimate the utility function of data over time, it is important to consider those portions of the value of data that vanish almost immediately and those that preserve or even gain value over time.

Fifth, data exhibits *economies of scale and scope*, but such effects are not identical to those exhibited by traditional economic goods. When it comes to data, economies of scale and scope go hand in hand, given the "quantum" properties of data (see above). The aggregation of data of the same type and for the same purpose generates enormous value, and the aggregation of different but complementary types of data for a single purpose can generate even more value. This, in turn, implies that those that can aggregate data from a variety of sources, especially if aided by powerful compute infrastructure and machine learning, can derive exponentially greater value from the whole data set. When this occurs, the aggregation of data leads to so-called big data computation, and the decoupling of value generation from value creation becomes potentially massive.

“

The value of data, from a societal perspective, is dependent both on the way in which data is collected, stored, processed and shared, as well as on the purpose for which data is used

Sixth, (digitised) data is increasingly *pervasive*. This trend became exponential since the Internet created an environment exclusively made of information, which gradually permeated the whole economy and society, leading towards a post-scarcity age.[4]

> In today’s “zettabyte” age, the generation, aggregation, transfer and re-use of data form a market, the sheer size of which approximately doubles every year, powers revolutions in many sectors and branches of human knowledge.

The constant, exponential increase in computational capacity (fuelled by Moore’s law) provides the necessary complement to AI systems that today reach trillions of parameters, crunching enormous amounts of data at the speed of light. In an end-to-end environment such as the Internet, the emergence of powerful network externalities and the rise of decentralised governance architectures further fuelled the explosion of data; and today, with the digital transformation of industry, the rise of cyber-physical objects promises an explosion of the number of connected devices, most of which will generate, receive and exchange data; and the rise of a world entirely made of data, from digital twins to the metaverse.

Seventh, (big) data is easily subject to *dual-use* considerations, in that the processing of large-scale datasets can lead to breakthroughs in science and knowledge but also to widespread surveillance of citizens and mental manipulation in the online environment. Thence, the value of data, from a societal perspective, is dependent both on the way in which data is collected, stored, processed and shared, as well as on the purpose for which data is used. Misuse or malicious uses of data can lead to bias, discrimination and security risks, whereas the pervasiveness of data can also lead to loss of human agency.

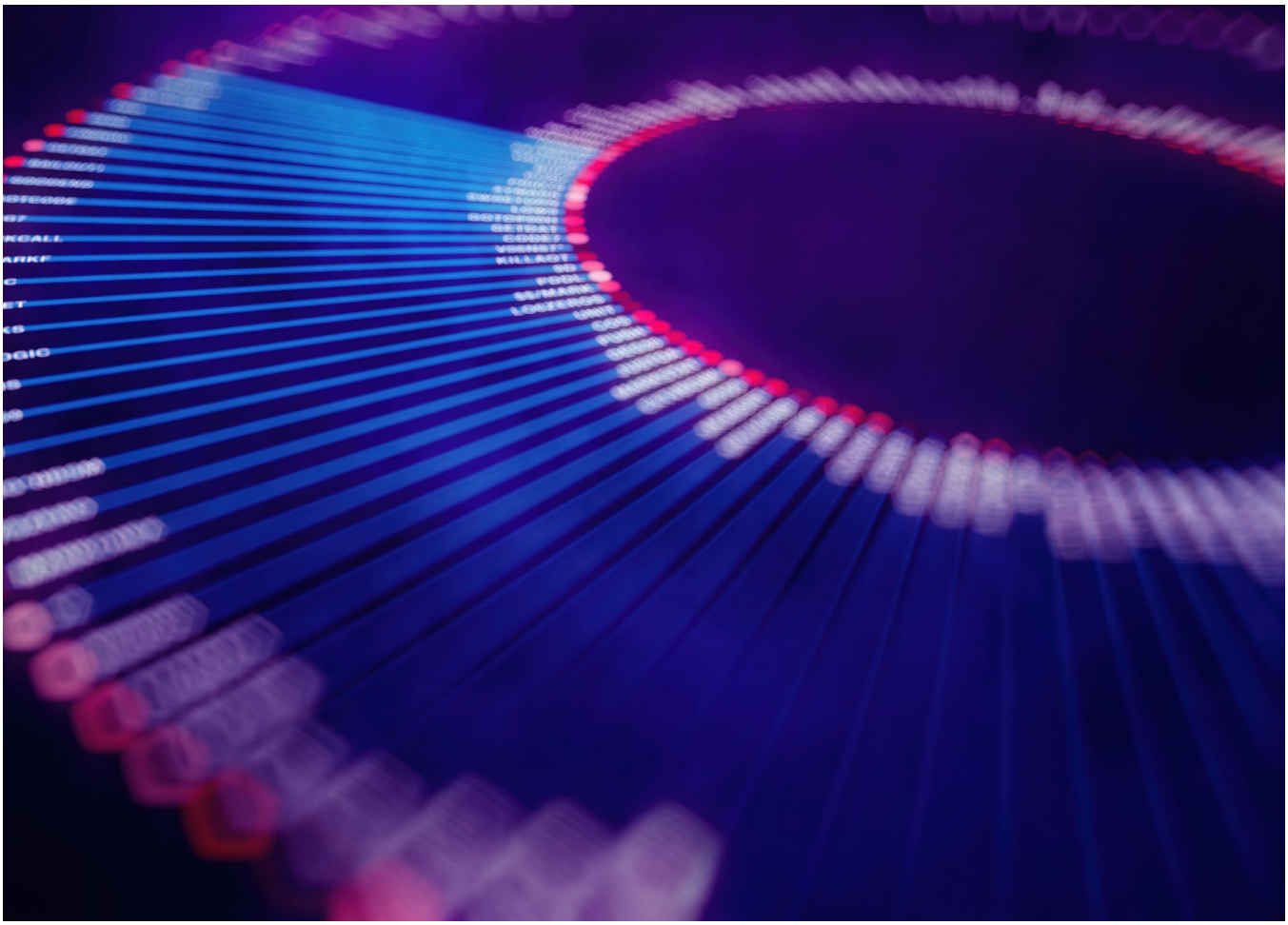
Eighth, the more it becomes widespread, aggregated, intermediated and processed through AI, the more data requires methods to ensure *verifiability and trust*. The need to ensure data verifiability and tracing has led to the emergence of a variety of technological solutions, including privacy- and confidentiality-enhancing technologies, and blockchain-enabled tamper-proof databases that can be used *i.a.* in proving data authenticity and, relatedly, in fraud prevention. The growing role of data in trade and supply chains also led to the need for trust-enhancing solutions, which reduce transaction costs (particularly information costs), thereby enabling efficient trading. At the global level, the need to ensure the “free flow of data with trust” found alignment in the context of the G7 under the initiative of the government of Japan.

> If seen from the perspective of diffusion, data is like water: it is globally (and increasingly) abundant but very often siloed and locally scarce.

Ninth, if seen from the perspective of diffusion, data is like water: it is globally (and increasingly) abundant but very often siloed and locally scarce. More specifically, in the current "zettabyte age" in cyberspace, data *is extremely concentrated and asymmetrically distributed*, to the extent that EU Commissioner Thierry Breton noted in 2020 that more than 90% of the data generated by European citizens and businesses is in the hands of a fistful of cloud-based (non-European) giants. As recognised i.a. by the European Commission in its impact assessment of the Data Act, while being potentially widely accessible, data is in practice subject to stunning asymmetries, to the extent that the economy is becoming polarised between data "haves" and "have-nots". Given its reliance on powerful data centres, high-capacity compute infrastructure and large-scale AI systems, the value of data can ultimately be reaped only (or mostly) by those very powerful players that dominate the internet economy. Hence the calls for a more equitable distribution of data, let alone its re-use for the general interest (see below, and also Stefaan Verhulst's companion paper).

Finally, one of the peculiarities of data is that it is *malleable* and *re-usable*. It can be decomposed, rebuilt and repackaged *ad libitum*, thus leading to endless possibilities for versioning, sampling, re-use, including through user-generated content, text and data mining and many other activities. As already mentioned, also thanks to the fact that digitised data exhibits near-zero marginal cost, the value of data normally incorporates an option value, which corresponds to the future uses that aggregating and elaborating a given data point could bring to those that can access and use it. Coyle et al. (2020, 2022) offer an interesting collection of methods and criteria that can be used to place a value on data, all depending on the main lens through which data is being observed.[5] They also distinguish between cost-based, income-based, and market-based methods to evaluate data.

Given these unique features, data escapes easy classifications and categorisations in social science and especially in economics. For example, the large literature on the valuation of personal data based on users' willingness to pay *(wtp)* for privacy largely misses the point, since users are cognitively unable to anticipate the value of multi-faceted, multi-status data to be aggregated by other entities. The situation that emerges is a de facto appropriation of data from unaware customers, who very often spontaneously communicate the data through social media and other online platforms. Contrary to what happens in the case of "takings" in public policy, where each owner's monopoly position warrants an expropriation *ex imperio* by public authorities, here a collective action problem prevents users from building a strong-enough bargaining power vis à vis large data aggregators, and ultimately leads to the capture of data and their related value by the aggregators themselves.

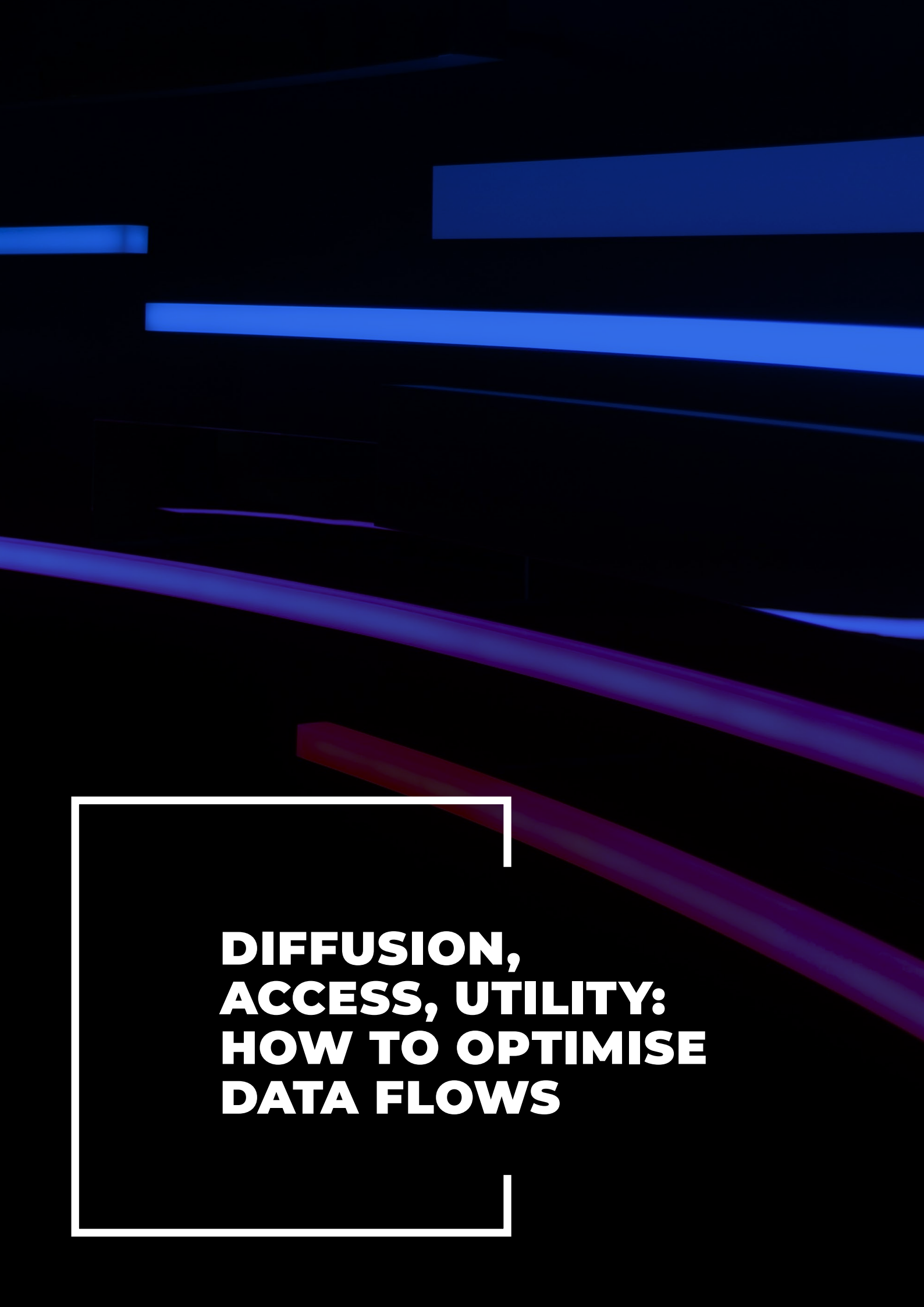


Moreover, expecting the remuneration of data to occur "at cost", or in any event the price of data to follow the underlying cost structure, makes little sense for several reasons: because the marginal cost of replicating data is often (very close to) zero; because many data-driven business models take the form of multi-sided platforms, where price tends to depart from cost and is rather a function of externalities; and also since there is no unique way of determining the cost of data, since the *wtp* for data by counterparties would depend on what that party plans to do with the data.

> Given these unique features, data escapes easy classifications and categorisations in social science and especially in economics.

Finally, betting on market forces to enable optimal data exchanges is often preposterous. "Coasian" solutions, in which the market leads to the redistribution of entitlements and thereby the attainment of allocative efficiency, are not practical in the world of data. The combination of informational asymmetries (Arrow's paradox) and the impossibility of comparing the *wtp* of potential counterparties makes transactions unlikely to occur.

The consequences for policymakers are far-reaching. They largely depend on the need to consider the peculiar economics of data and apply it to a wide variety of applications and use cases, in which defining the optimal circulation of data is not as straightforward as economists and policymakers traditionally thought. Section 1 below expands on this specific issue by looking at the utility function(s) of data. Section 2 draws conclusions as regards optimising policies for today's massive data flows. The concluding section provides some perspectives on future research.



DIFFUSION, ACCESS, UTILITY: HOW TO OPTIMISE DATA FLOWS

DIFFUSION, ACCESS, UTILITY: HOW TO OPTIMISE DATA FLOWS

One of the most confusing aspects of data is its utility function, *i.e.* the evolution of its value as a function of its diffusion. On this aspect, many economists have dubbed data a "public good", which suggests that the greater a given data's diffusion and the possibility of access by third parties, the greater its value. Just as accessing the signal sent by a lighthouse or a radar has the same value for a sailor, irrespective of the number of sailors that observe them at any given moment of time, the value of data would increase alongside its diffusion, with no rivalry in consumption, and no (need for) exclusivity in access. In other words, if one takes this perspective, data exhibits the same non-rivalry and non-exclusivity features of public goods in economics, and as such, its diffusion and access should be incentivised as much as possible. The consequence of this vision would be that policymakers willing to maximise the value of data should also facilitate its flows as much as possible. From this viewpoint, the emergence of the World Wide Web in its original design, an end-to-end "network of networks" with little or no filtering of digital data flows, created the perfect preconditions for the diffusion of digitised data and, as such, would seem to represent the perfect solution for data policy.

> A "let the data flow" or *laisser-partager* approach has indeed characterised the first three decades of cyber policy, in which online intermediaries have been shielded from any responsibility to control and filter the data flowing on their servers and platforms.

At the international level, initiatives such as the G7 "free flow of data (with trust)" are also inspired by the same approach.

As a matter of fact, the public good nature of some types of data is undeniable, at least when one observes this phenomenon from a static viewpoint. Scientific breakthroughs, such as the discovery of new, powerful drugs for existing diseases, should, in principle, be shared as widely as possible. When he refused to patent the vaccine for poliomyelitis, Albert Bruce Sabin motivated his choice with the need to make the solution as widely and readily available as possible.[6] Likewise, open access data is considered to be a key element of scientific research, just as the training data and algorithmic code used to develop powerful AI-enabled solutions (e.g. AlphaFold). Several programmers, institutions and corporations can, in principle, use the same code to develop their own solutions with no rivalry effects. Once the code has been released in open format, theoretically, no one can exclude any others from using it. Much in the same vein, the release of code for powerful large language models, or generative AI models, enables innovation by allowing researchers to access key data and develop derivative solutions by adding to the code or choosing specific training data and use cases. The best circulation model for data as a public good would thus be a "zero-cost liability rule" or at least a low-cost compulsory access regime.

At the same time, there are reasons to believe that characterising data as a public good does not tell the whole story of data's utility function. For example, when seen from a dynamic perspective, the cases mentioned above would look quite different. The absence of prospects for exclusive exploitation of scientific results would most likely determine a weakening of incentives for scientific discovery and R&D investment in the first place. In other words, as widely acknowledged, non-rivalry creates a lack of incentives to produce valuable data in the first place. Moreover, in many cases, access to and consumption of data is rivalrous, and competing

> Some types of data feature public goods characteristics and, as such, see their value maximised alongside their diffusion. (...) some data is worth the most when kept private.

access to data creates congestion: this is why several scholars have dubbed data a "common good", as such subject to the well-known "tragedy of the commons" in the absence of well-specified property rights.

Most importantly, not all data reaches the peak of its value when openly accessed and shared. Consider the following examples:

- Bob discovers that a world champion in boxing has secretly agreed to lose the next match, where he is the obvious favourite; he then decides to massively bet on the victory of his rival. The value of that information reaches a peak when kept private and rapidly falls to zero as the information is shared ahead of the match.
- Alice is the CEO of a corporation which plans to launch a hostile takeover of another business through a swift and coordinated plan, which requires the help and support of other investors, such as hedge funds. If she keeps the information for herself, that information is worth nothing, as the plan would not be executable. If she shares the information openly, that information would be worth very little as the market and the target company will anticipate the move. That information reaches a peak of its value only when shared among a limited group of individuals, and as such, its utility function takes an inverted U-shape.

In summary, some types of data feature public goods characteristics and, as such, see their value maximised alongside their diffusion. I will call these data "Type A". Other data reach a peak of their value when shared within a contained group. I will call these data "Type B". Lastly, some data is worth the most when kept private. I will call these data "Type C". Figure 1 below shows the three types of data and the related utility functions.

Figure 1. Types of data, social utility and diffusion

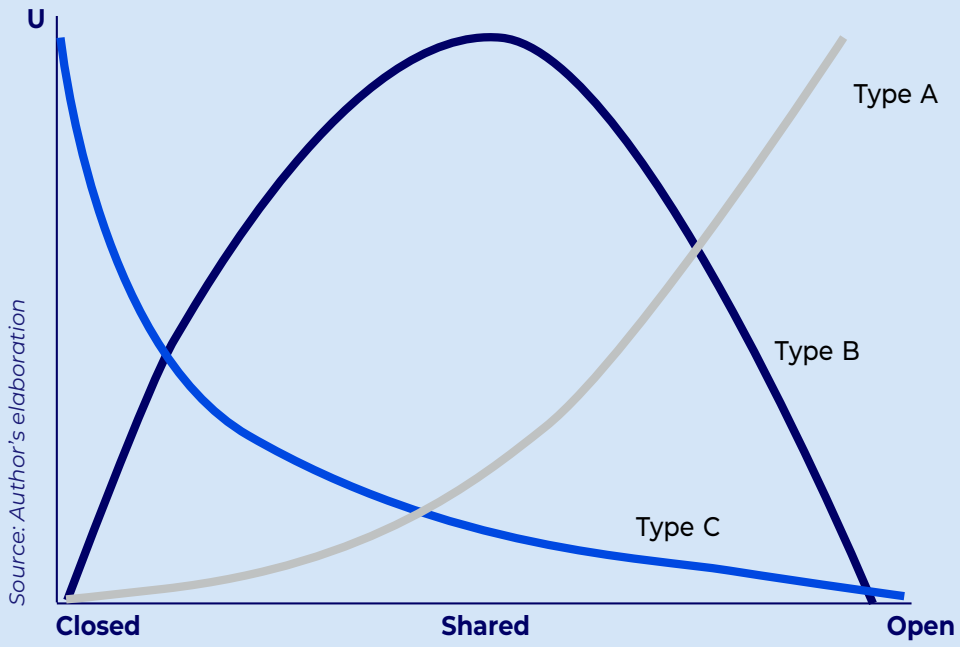
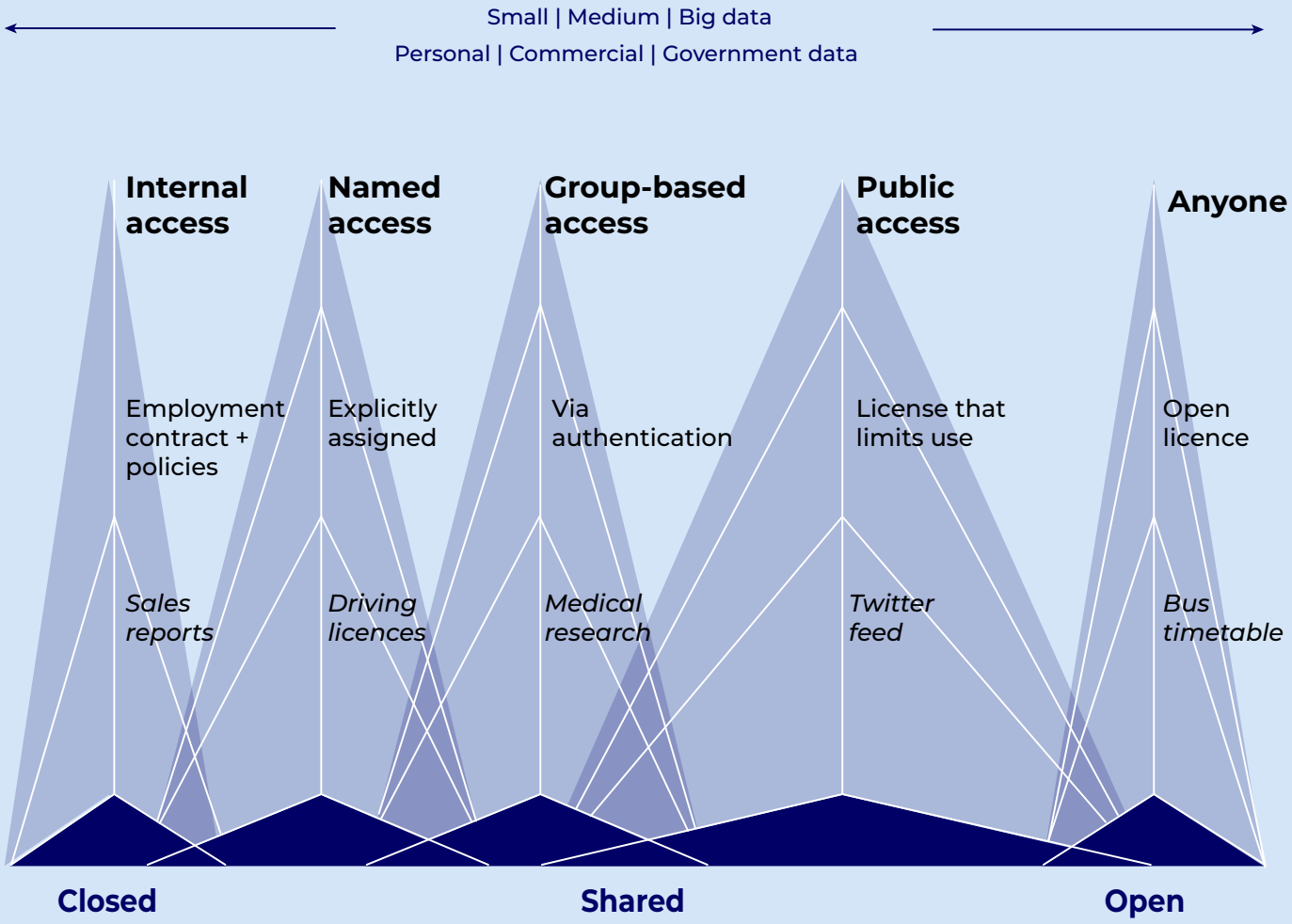


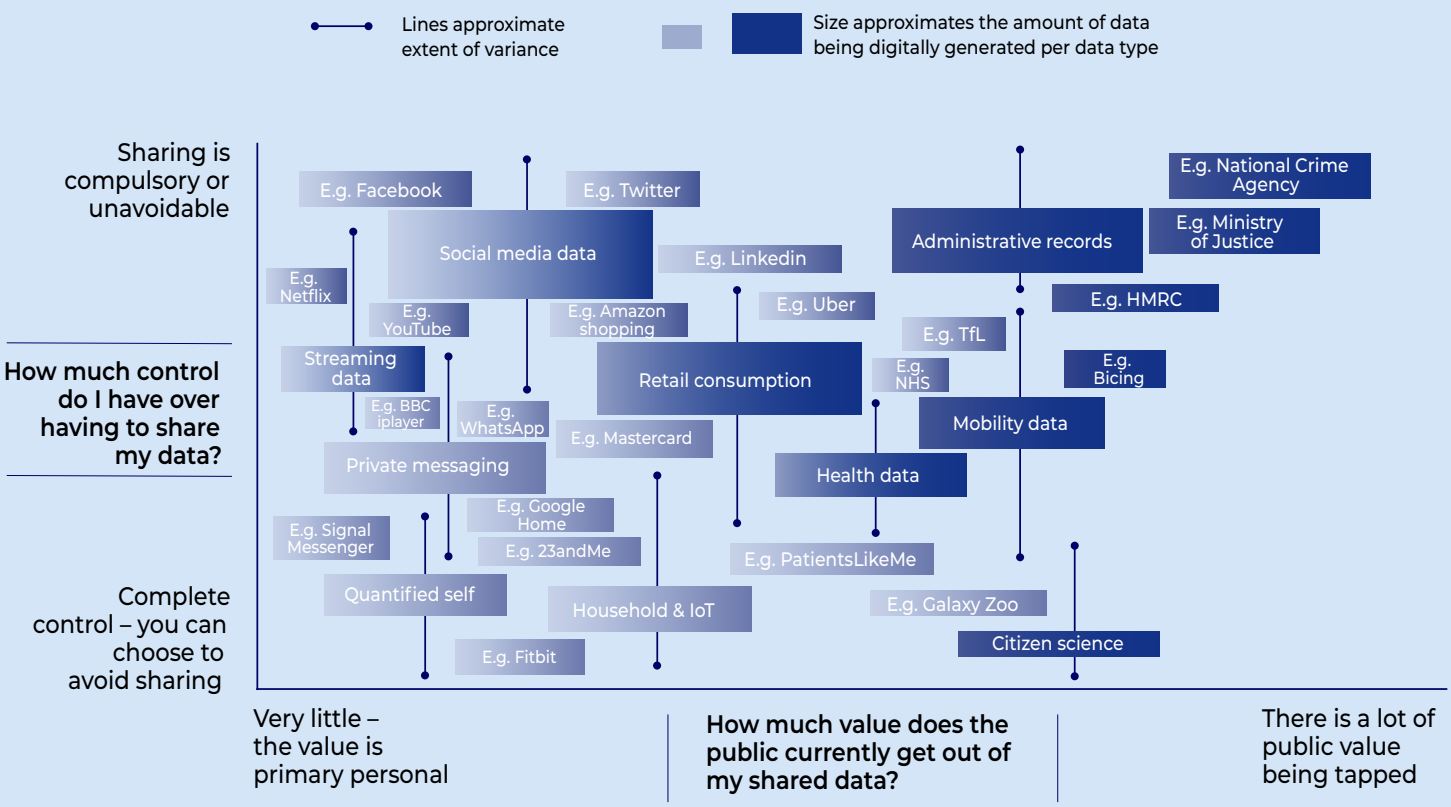
Figure 2. Data diffusion, mode of circulation and access



Source: Coyle et al. (2020)

Coyle et al. (2020) offer a more granular analysis of the relationship between the diffusion of specific types of data and their utility, coupled with the mode of circulation and diffusion of each type. [7] As seen above in Figure 2, data used in the context of business and trade (such as sales reports) tends to remain "closed" and reach maximum value when kept private. It is normally generated and circulated through restrictive contractual policies and remains confined in an internal access mode. Other types of data, such as medical research, are issued and retrieved via authentication and end up being shared across professionals (as in the emerging EU health data space) for health-related purposes but have more value for society when they are not fully and openly shared. A bus timetable, on the contrary, features the characteristics of a public good and is normally made available to anyone on a non-rival and non-exclusive basis.

Figure 3. The current landscape of data governance



Source: Mulgan and Straub (2019)

Similarly, already in 2019, Geoff Mulgan and Wolfgang Straub classified data along two slightly different axes; the so-called "public value" of data (i.e. the extent to which the public gains if the data is shared), which could be equated to its social utility; and the level of control of the data holder over having to share data, which is tantamount to the access and circulation regime mentioned above.[8]

Ideally, when looking at Figure 1, an optimally calibrated data policy should ensure that Type A data gets shared as much as possible; Type B data is shared to the extent that it maximises its value (the actual peak may correspond to varying levels of diffusion); and Type C data is kept private.

There is room for a data governance strategy that tries to optimise data flows by removing the distortions, asymmetries and power concentration effects that characterise the current digital ecosystem.

But how can the policymaker anticipate the optimal level of diffusion for each type of data?

> In an ideal ("Coasian") world, the best mechanism for optimising the level of data diffusion would be the market: regardless of the constraints or facilitating provisions included in legislation, if transaction costs are low, the parties would be able to transact over data, in a way that ultimately achieves allocative efficiency.

In line with the Coase theorem, this means that even if rules such as GDPR restrict the use of personal data by placing a veto power ("property rule") in the hands of the data subject (in the form of informed consent), the data subject will end up negotiating away such property rule in exchange for sufficient compensation. Likewise, data that the legislator placed in Type C (by prescribing its free flow) could be given a more restricted flow if a group of stakeholders decided to contract into a joint ownership or management scheme.

However, as already explained in the previous section, the economics of data is fraught with transaction costs, further exacerbated by the "quantum" characteristics of data, which in turn make it difficult to complete transactions on the basis of a homogeneous wtp, as well as by the behavioural and informational features of data, which make it difficult for individuals to place a value on their personal data. In cyberspace, this has led to an over-circulation and systematic re-use of data, regardless of their public and private value, as well as an under-diffusion of data that produces more value when shared. As this reduces the overall value of data, as well as the utility that the public derives from it, there is room for a data governance strategy that tries to optimise data flows by removing the distortions, asymmetries and power concentration effects that characterise the current digital ecosystem.

DESIGNING AN EFFICIENT DATA STRATEGY: A DECALOGUE FOR POLICYMAKERS

DESIGNING AN EFFICIENT DATA STRATEGY: A DECALOGUE FOR POLICYMAKERS

Given the unique complexity of the data landscape, regulators around the world have gradually realised that a simple *laissez-faire* approach would not be in line with the public interest.

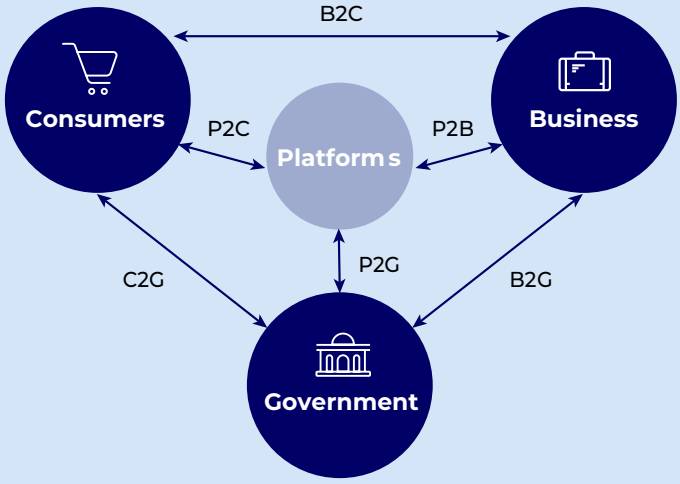
> In the case of the European Union, the need to act to promote an optimal diffusion of data was gradually coupled with industrial policy stances due to the fact that data were increasingly concentrated in the hands of few, non-European players.

While this could be an unfortunate and undesirable fact when it came to personal data, for industrial data, this would become an existential threat to Europe's businesses. The rationale is simple: most of the value of industrial products increasingly resides in the data layer and in the value-added services it enables; the digital transformation of industry, with the rise of industrial AI and Internet of Things (the so-called Industry 4.0), is leading towards a massive generation of industrial data, with a projected one trillion connected devices by 2035; these data, if current trends continue, will land in the hands of those same non-European cloud giants that currently dominate the market; as a result, Europe would lose most of the value generated by its industrial production.[9]

Not surprisingly, then, it was the European Union that took the first initiative to define a comprehensive data strategy, presented in February 2020 and later implemented with a series of landmark initiatives such as the Data Act and the Data Governance Act, but also the launch of sectoral data spaces (for now, in health) and the attempt to create a federated cloud environment through the GAIA-X project. In line with what is described above, the logic behind Europe's activism in this domain is both an industrial policy and redistribute the value of personal and industrial data, opening the market to competition. The European Commission, in assessing the prospective impacts of the Data Act, identified many problems to be tackled in order to turn the tide of the first three decades of the Web. These included *i.a.*, the limited ability for consumers to realise the value generated by their use of the products, the low levels of data availability for creating added value in B2B relations, and inefficient practices for the use of private sector data by the public sector.

More generally, it is useful to map the existing challenges in data governance by identifying (potential) data flows between all stakeholders in the data ecosystem, as shown in Figure 4 below.

Figure 4. Data flows between Consumers, Businesses, Platforms, and Governments



Source: Author's elaboration

The difficulty for policymakers is that not all the flows shown in Figure 4 must be encouraged. Rather, the following goals are emerging (particularly in the EU and China) as important for policymakers in the data space:

GOAL 1.
Minimise the flow of personally identifiable data.

The need to protect privacy by limiting the non-exclusive characteristics of personally identifiable data is widely recognised around the world. However, as mentioned above, the need for proactive intervention to sanction the appropriation and re-use of personal data (e.g. through the EU GDPR) is linked to a "property rule" treatment of the right to privacy, which presupposes that individuals can properly assess the value of their personal data, and express a willingness to pay for privacy preservation.[10] As this appears to be an acrobatic assumption due to behavioural biases and the specific features of data as an economic good, the effectiveness of these interventions to date has been quite limited. In the future, the technological protection of privacy (e.g. through privacy-enhancing technologies, or PETs) appears to be a more promising avenue.

GOAL 2.
Encourage businesses to share data (altruism).

Around the world, the need to facilitate the sharing of certain data is frustrated by the lack of "clear rules and processes" in place that address the issue of data altruism. Facilitating measures, as mentioned by the European Commission already in 2020, "would ensure that more data becomes available for the common good, and would increase trust in altruism schemes".[11]

GOAL 3.
Enable managed data-sharing within and across data spaces.

For so-called "type B" data, it is important to ensure that data can be pooled and co-managed while at the same time avoiding that it can flow freely and is thereby appropriated by more powerful data aggregators. The need to promote joint management of data in specific sectors or for specific purposes has given rise to the idea of so-called "data spaces", initially by industry players (e.g. in Germany) and later by policymakers. Data spaces pursue at once the objective of promoting a more equitable distribution of data along the value chain (e.g. avoiding value capture), the optimisation of services by enabling coordination between different stakeholders in a given ecosystem, and the aggregation of structured and unstructured data from various sources to enable the provision of services for the public good.

GOAL 4.
Avoid data hoarding and value capture.

The need to avoid consumers being locked in by multi-sided platforms that bundle devices with data, which platforms can exploit in various ways (including excluding competitors from access to such data by invoking data protection rules as a "shield"), has given rise to proposed measured aimed at unbundling data from devices. In the EU, such measures are included in the Data Act but also reinforced by measures imposed on so-called "gatekeepers" in the Digital Markets Act. Besides, policymakers are gradually looking at ways to encourage efficient B2B data sharing by promoting far contractual conditions in data sharing contracts (see Goal 7 below).

GOAL 5.
Facilitate switching between cloud and edge services.

The crystallisation of market power in the hands of a few cloud-based giants, and the consequent lack of choice for consumers, is being addressed (in addition to measures already mentioned above under goal 4) also through mandatory interoperability requirements aimed at helping new entrants access the data they need to viably compete with large-scale incumbents. In the European Union, the GAIA-X project envisages the creation of a federated cloud infrastructure, which entails *i.a.* the imposition of such interoperability obligations to create market opportunities for smaller-scale European cloud operators; in specific sectors, data-sharing with competitors/new entrants is also made explicit, even if regulators have struggled to introduce well-balanced frameworks providing incentives and compensation for players having to share data (see, for example, the EU Second Payment Services Directive, or PSD2). However, the problem is that as users consume more sophisticated services, and not just infrastructure, consumers normally do not face a wide choice of an alternative providers offering equivalent services.

GOAL 6.
Oblige or incentivise businesses and platforms to share data "for good" and emergencies.

Governments are increasingly realising the usefulness of using big data "for good", i.e. to deliver value-added services in the interest of society. However, data are very often in the possession of the private sector, and even when they are held by public authorities, they often feature different levels of quality and very low interoperability. The experience of COVID-19, with governments having to negotiate access to privately held data in order to better track mobility and the possible spread of the virus, further inspired the adoption of rules or self-regulatory schemes aimed at enabling B2G data sharing by protecting data with liability rules, rather than property rules; and by remunerating data at cost (whatever this means in practice). In practice, these solutions are still in their infancy and await both adequate definitions (which authorities, which data, in what format, what is the public interest, etc.) and the fine-tuning of sharing conditions, including a reference formula for assessing compensation for data access. At the same time, the possibility for public authorities to release open data for private players to develop value-added services (so-called G2B sharing) should be given further attention in the public debate: currently, several governments operate open data platforms, extreme cases being so-called "Government as a Platform" (GaaP) solutions such as Estonia's X-Road.

GOAL 7.
Ensure fair contractual conditions in data-sharing contracts.

Imbalances in bargaining power and high transaction costs between (large-scale) platforms and businesses (P2B), as well as between businesses of different sizes and power (B2B), are leading policymakers to adopt specific initiatives to ensure that data sharing takes place when efficient. At the EU level, both an ad hoc P2B regulation and specific provisions in the Data Act.

GOAL 8.
Increase data sovereignty.

In recognising the enormous value of data for the economy and society, many governments around the world are pursuing enhanced data sovereignty, a term which can be interpreted in two ways:

(i) from a geo-economic perspective, as the need to retain data in the territory of the country, as well as in the hands of domestic players (see, e.g. the EU Certification Scheme for Cloud Services); and

(ii) from a micro perspective, as the need to ensure end-user control over the diffusion of data.

This latter goal is increasingly felt as essential for policymakers and civil society around the world, and relies on technological solutions such as so-called Personal Information Management Systems (PIMS), such as those developed under MyData, IHAN or Tim Berners-Lee's Solid project. At the same time, sovereign data solutions do not always come without costs. First, having data localised in one's territory does not automatically make it more secure against unauthorised access or even loss, nor does it really solve the concentration risk. Second, data localisation requirements could deprive consumers of sufficient choice among competing offers and thereby lead to higher prices and lower quality.

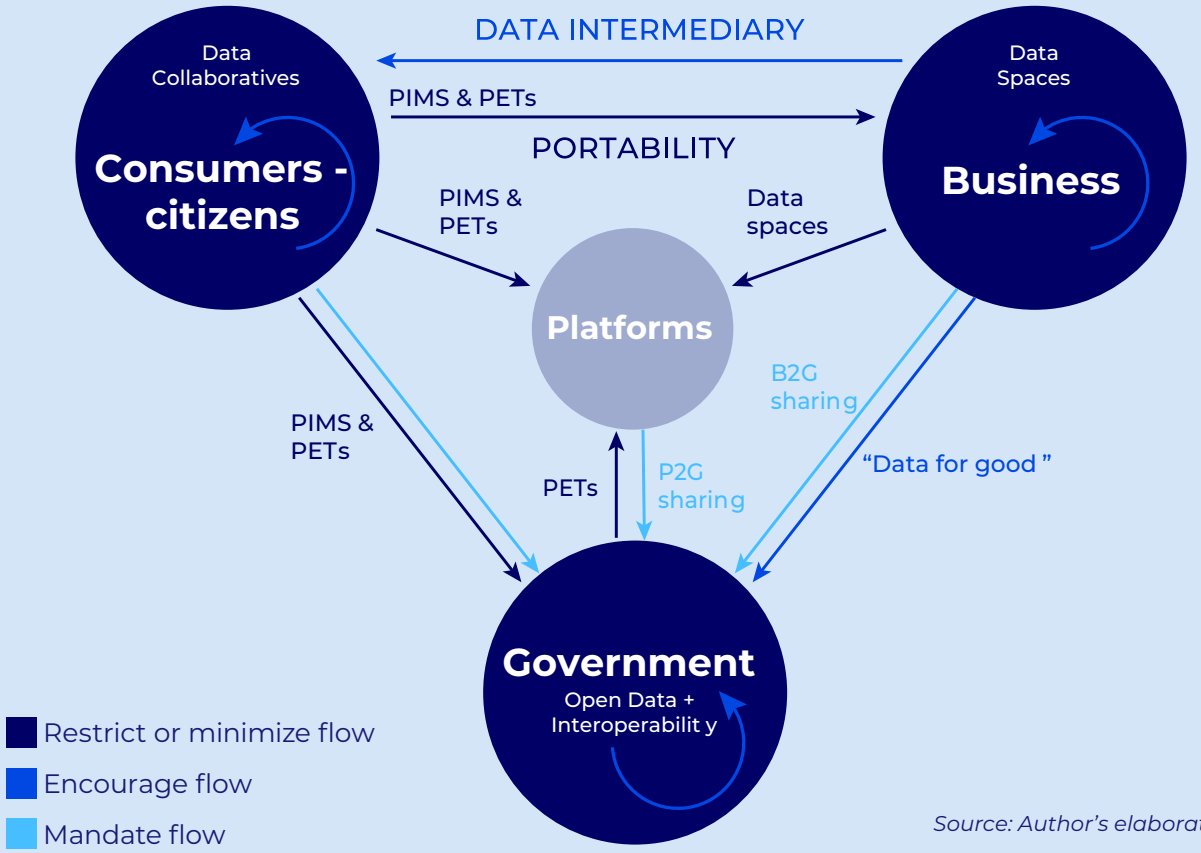
GOAL 9.
Create trusted and independent data intermediaries.

Any attempt to facilitate the flow of data between consumers/citizens, businesses/platforms and governments is doomed to fail unless these entities establish mechanisms to enhance trust between them. Transaction costs can be reduced in various ways, including through technology (e.g. PETs, blockchain), yet the availability of independent, trusted intermediaries is likely to be the best way to achieve optimal data flows in the medium term. The most important features of a data intermediary are skills (see goal 10 below), accreditation and independence, intended as the fact that the intermediary is not using the data it collects as part of a multi-sided business model, in which data are re-used without the meaningful and informed consent of the end user. In the European Union, a specific piece of legislation, the Data Governance Act, was enacted to create the preconditions for the emergence of data intermediaries; the latter are also facilitated by provisions in the Data Act, the Digital Markets Act, and technical specifications in GAIA-X and in industrial data spaces.

GOAL 10.
Promote data stewardship and literacy.

In order to facilitate collaboration over data, it is essential that consumers, businesses, governments and intermediaries develop sufficient skills. Demand for data-related services has led to the rise of a new profession, which Verhulst and Young (2022) define as "data steward".[12] The skills that the steward should have include Data Audit, Assessment & Governance; stewarding partners: partnership and community engagement; internal coordination and data operations; nurturing and sustaining data collaboratives; and disseminating outcomes and communicating activities.

Figure 5. Restricting, encouraging and mandating data flows



Source: Author's elaboration

> Eventually, the role of the policymaker is to create the governance arrangements and the policy preconditions that would lead to optimal data flows, which can nurture and feed the economy and society, leading to enhanced prosperity.

Figure 5 above summarises the types of data flows that policymakers typically try to restrict (as they pertain mostly to Type A or Type B data); data flows that should be encouraged; and data flows that could be made mandatory under specific conditions, as they would not otherwise occur. The figure also shows some of the key technologies and governance arrangements that can facilitate the optimisation of data flows: personal information management systems (PIMS) and privacy-enhancing technologies (PETs); data collaboratives, especially among citizens; data spaces for businesses; and interoperability arrangements across and between governments.

CONCLUSION: DATA AS THE “FIFTH ELEMENT”

CONCLUSION: DATA AS THE "FIFTH ELEMENT"

Economists have struggled, and continue to struggle, to capture data's peculiar, elusive dynamics. Given the growing pervasiveness of the data economy, developing a more granular understanding of how policymaking could help optimise data flows is urgent. This could offer governments a key toolkit to ensure that the digital transformation benefits society as a whole and that data is put to its best possible uses. At the same time, crafting an effective, efficient, human-centric and sustainable data policy is not easy, especially since the Internet has been so far a largely unregulated space, where data could flow freely and with very limited control by both public and private entities.

In the coming years, research will be able to contribute extensively to optimal data policy by filling a number of outstanding gaps. There is a strong need for deeper research on data valuation methods for the purposes of B2G and B2B data sharing: at the moment, for example, the solutions outlined in the EU Data Act fall short of providing sufficient regulatory certainty for market players wishing to engage in data altruism and sharing, as well as for actors facing obligations to share data with public authorities. Moreover, there is a need for deeper studies on how to keep track of data flows and use data for training purposes, especially for the development of generative AI systems and for data flows between connected objects (and related smart contracts): as the world ushers into an age of low-value, high-frequency transactions, researchers must support policymakers wishing to provide a suitable legal and regulatory system for future, dense and immersive data flows.

Furthermore, the human component is increasingly essential when it comes to managing data. Yet governments and the private sector seem to lack basic data stewardship skills, which must be further analysed and translated into educational programmes.

> In this respect, it is of utmost importance that data governance becomes the subject of a new social contract, in which citizens and civil society are aware of the importance of optimising data flows and retaining control rights of the diffusion of certain types of data (self-sovereignty); and at the same time, act as key pillars in the monitoring of how data is used by the private sector, and for the public interest.

There is no overstating the importance of data for the future of our societies. In a world dominated by connected objects and immersive AI systems, data represents a true "fifth element" in addition to fire, air, earth, and water. This is why social scientists must develop a "new physics" for the data-immersive economy and society. This paper tries to stimulate reflection and further research in this extremely crucial domain.

ENDNOTES

1 Thouvenin, F., & Tamò-Larrieux, A. (2021). Data Ownership and Data Access Rights: Meaningful Tools for Promoting the European Digital Single Market? In M. Burri (Ed.), Big Data and Global Trade Law (pp. 316-339). Cambridge: Cambridge University Press. *https://doi.org/10.1017/9781108919234.020*

2 Arrow, Kenneth J. (1962), Economic Welfare and the Allocation of Resources for Invention, in The Rate and Direction of Inventive Activity, 609 (National Bureau of Economic Research ed. 1962).

3 Mazzucato, M. (2018), The Value of Everything: Making and Taking in the Global Economy, New York: Public Affairs, 2018; UNCTAD (2019), Value Creation and Capture: Implications for Developing Countries, Digital Economy Report 2019.

4 Desai, Deven R. and Lemley, Mark A., Scarcity, Regulation, and the Abundance Society (June 30, 2022). Stanford Law and Economics Olin Working Paper No. 572, *http://dx.doi.org/10.2139/ssrn.4150871*

5 Coyle, D., Diepeveen, S., and Wdowin, J. (2020). The value of data summary report. The Bennett Institute for Public Policy, Cambridge; Coyle, D. and A. Manley (2022), What is the Value of Data? A review of empirical methods, Policy Brief, Cambridge University.

6 Orsini D., M. Martin (2022), Albert Bruce Sabin: The Man Who Made the Oral Polio Vaccine. Emerg Infect Dis. 2022 Mar; 28(3):743–6.

7 Coyle, D., Diepeveen, S., and Wdowin, J. (2020). The value of data summary report. The Bennett Institute for Public Policy, Cambridge.

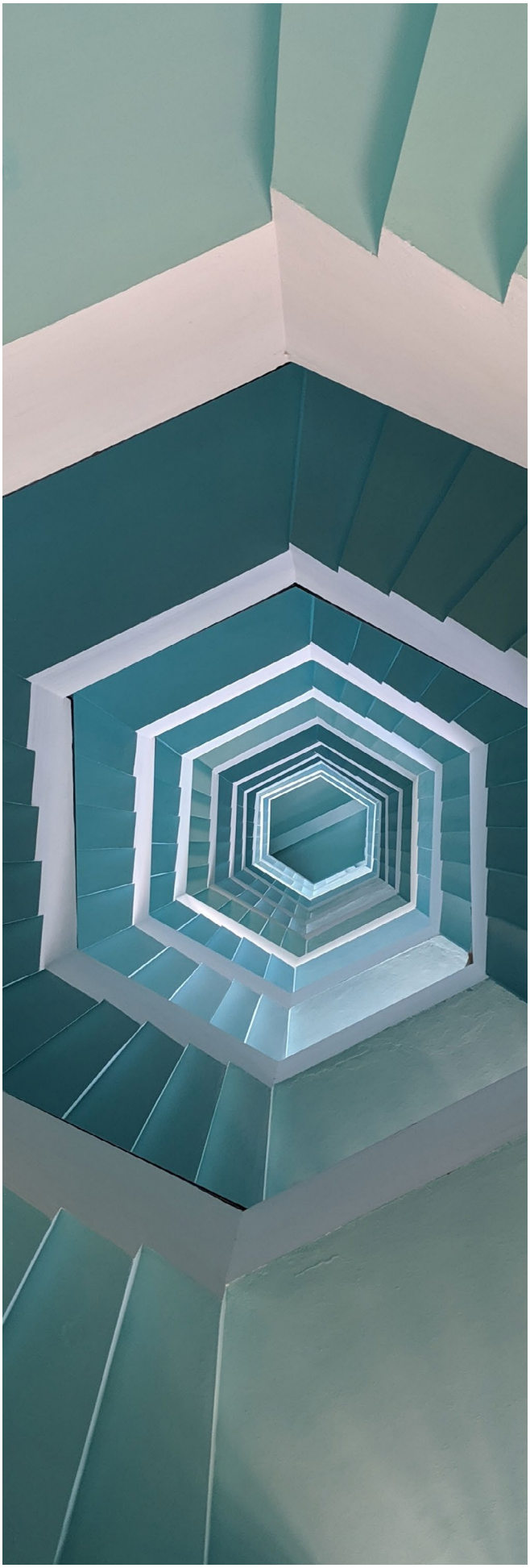
8 Mulgan, G. and V. Straub (2019), The new ecosystem of trust, Nesta, UK

9 Renda, A. (2020) Making the digital economy "fit for Europe". European Law Journal. 2020; 26(5–6): 345– 354. At *https://doi.org/10.1111/eulj.12388*

10 Skatova A., R. McDonald, S. Ma, C. Maple (2023). Unpacking privacy: Valuation of personal data protection. PLoS One. 2023.

11 European Commission (2020), Impact Assessment report accompanying the Proposal for a Regulation on Data Governance, SWD(2020) 295 final, Brussels, 25.11.2020.

12 Verhulst, S. & A. Young. (2022). Identifying and addressing data asymmetries so as to enable (better) science. Frontiers in Big Data. 5. 10.3389/fdata.2022.888384.



ie UNIVERSITY

CENTER FOR THE GOVERNANCE OF CHANGE

AUTHOR:

Andrea Renda
IE University

RECOMMENDED CITATION:

Renda, A., "Data Policy: A Conceptual Framework,"
IE CGC, June 2023

Design: epoqstudio.com

cgc.ie.edu