1

Evaluation as Power

In 1951, Frédéric and his wife Irène were invited to grace the Congress of Polish Science in Warsaw with their presence, as the members of the French delegation. They were among the key officials who had come from abroad. He was the first president of the World Federation of Scientific Workers – the group of scientists who supported communism – and with his wife, had been awarded the Nobel Prize in Chemistry in 1935. Irène Joliot-Curie was the daughter of Maria Skłodowska-Curie and Pierre Curie, who had also been the recipients of the Nobel Prize.

During Stalinism's apogee, the Polish Communist Party decided to reorganize the country's entire scientific landscape and to subordinate all research institutions to the Party. The key means by which they sought to achieve this was by establishing the Polish Academy of Sciences based on the Soviet model. It was during the congress that the framework of this new state institution was presented. The plan was that the Academy replace all learned societies and ministerial institutes as well as most of the research projects conducted at the universities. As the key political institution in science sector, it had to contribute to the Six-Year Plan that was concentrated on increasing the heavy industry sector.

Frédéric was one of the keynote foreign speakers, and he spoke just after the members of the Russian Academy of Sciences. They were all, however, preceded by a speech by a miner and officially recognized "model worker," Alojzy Mozdrzeń, who welcomed the Polish and foreign scientists on behalf of working people. This Polish miner assured the audience that all workers understood that the development of science is based on changing science's program so that it serves the people. Starting with the congress, all research efforts within the Academy had to be planned and coordinated in such a way as to contribute to the economy and – in the case of the humanities and social sciences – to a national culture.

Thus, the Communist Party decided to use its power to reorganize science in Poland, bringing it under its full control. In 1951, the Polish parliament passed the act that established the Polish Academy of Sciences, and one of its titles was

devoted to the "planning of research and reporting." From that moment, research projects could be conducted only if they were in line with the Party's plans and implemented by the Academy. Ruled by the Communist Party, the state assumed the control, governance, and evaluation of all research, opting to fund only those fields that contributed to the economy or were in line with the idea of "Soviet man." In this way, a national ex ante evaluation of research was established for all fields. A systematic evaluation of research was carried out, and the key criterion was whether a given institution, group of researchers, or single researcher had contributed to the plan.

Using its power to redefine the meaning of all scientific endeavors, the state thereby introduced new technologies of power through the central planning of research and reporting. These changes transformed science in Poland, as in other countries of the Eastern Bloc, and their effects were evident for many years to come.

* * *

During the Cold War, on each side of the Iron Curtain, a distinct way of thinking about science and its role existed.

The West perceived science mostly as "pure science," that is as an autonomous field organized by an ethos that defended its autonomy against outside influence (Merton, 1973). In 1962, Michael Polanyi characterized science as the republic in which scientists could freely select scientific problems and pursue them in light of their own personal judgments (Polanyi, 1962). Moreover, in this republic, the work of researchers was assessed according to its scientific value that was defined by its accuracy, systematic importance, and the intrinsic interest of its subject matter. In the republic of science, scientists were expected to keep a distance from public affairs.

The (socialist or communist) East, by contrast, perceived science through its social function. John D. Bernal (1939), an Irish scientist who politically endorsed communism, published *The Social Function of Science* in which he presented the idea of science as a tool for supporting a centrally planned society and industry. He considered that the way that science was organized in the Soviet Union was the best model for serving the people, nations, and societal needs. On the Eastern side of the Iron Curtain, the primary function of science was not therefore to cultivate a "pure science" oriented toward solving puzzles but rather to serve socialist society and its economy. Given that it was necessary to reorganize the whole scientific landscape in Poland in order for this to happen, the process was centrally planned and implemented. Science was therefore to serve rather than to build and sustain its own realm.

Today the Iron Curtain is a thing of the past. And yet I agree with Roger Pielke Jr. who, in assessing Bernal's legacy, states that his "ideas on the social function of

science have triumphed on nearly every count" (Pielke, 2014, p. 428). While neither the Soviet Union nor the Eastern Bloc exist anymore, at a global level, science is perceived through its social function. For instance, for several years now, one of the key criteria in Western research evaluation systems in the UK, Australia, or the Netherlands is the societal impact of research, which is defined mostly as the effects of research on the economy, culture, society, or quality of life beyond academia (De Jong & Muhonen, 2018; Derrick, 2018). During the Cold War, scientists in the West believed that they had to keep their distance from public affairs. Now, by contrast, they have to contribute to them and to solve problems, which today are called *grand challenges* (Omenn, 2006).

The current system of science funding is hybrid. On the one side, there are competitive grants and block funding for science and higher education institutions. On the other side, various institutions run programs by which specific *societal* challenges like pandemic research related to COVID-19, well-being, food security, and resource efficiency are defined as the goals of research which might be eligible for additional funding. This applies to institutions around the world – from the National Institutes of Health in the US to various agencies in European countries, to China and countries in all other regions. Moreover, global actors like the European Commission also fund research within the lens of grand or societal challenges.

One can observe that while the language with which the role of science is described has changed, Bernal's ideas are still vital. In the descriptions of grand challenges, it is difficult to find phrasing about research needing to serve the people, but one does find that such policy priorities address major concerns shared by citizens. Nonetheless, researchers can still apply for various grants in line with their own research interests, as members of Polanyi's republic of science. At the same time, however, a parallel funding path is becoming increasingly important: that of solving grand and societal challenges through centrally planned research subjects and themes. In either case, the societal impact of research is becoming as important as the scientific value of an investigation's results.

Within these two models, states have played an important role as actors that could either guarantee the autonomy of the republic of science or require that research has a societal (often predefined) impact. For states to achieve their interests, technologies of power are necessary, and one of the key technologies is evaluation.

1.1 Technologies of Power

Evaluation as a social phenomenon drives today's society. It serves to determine the worth, merit, or usefulness of something by reference to a set of values and varied protocols, instruments, and goals that are external both to evaluation itself and to what is evaluated. Dahler-Larsen argues that we actually live in an "evaluation

society" (2012) in which evaluation not only describes what has worth or represents value from the evaluation perspective but also describes and constitutes what evaluation claims to measure (Dahler-Larsen, 2015). This observation suggests that in the evaluation process, one does not observe a static relation between those who evaluate and that which is evaluated. Evaluation is rather a dynamic social process of constructing the evaluated objects in and through evaluation itself. In this book, I build on Dahler-Larsen's conception of the constitutive nature of all evaluation processes. This nature implies that evaluation is perceived and used not only as a tool for determining value or merit but also as an instrument of social change across various policy regimes.

Evaluative power is the capacity of the state (or other actors like global organizations) to influence and shape the definition of a key area and to change the behaviors and practices of individuals or institutions by deploying varied technologies. I use this term to name power relations produced by the state and its various policy instruments or, in other words, technologies of power in the science sector. The state's capacity to influence individuals and institutions is mediated by power relations. Evaluative power in science is based on constructing measures and measuring science and research. A key technology that serves to produce and sustain the evaluative power of the state in science is the research evaluation system.

A technology of power is a medium by which the state realizes its interests (e.g., priorities in some research areas) as the owner of public funds. A technology of power is embodied as a set of protocols, metrics, indicators, and policy aims. Like other media, such an embodied medium is not neutral (McLuhan, 1994). This means that through the state's process of constituting this technology, the public sector is influenced, shaped, and potentially transformed. A research evaluation system as a technology of power might also be understood as an "evaluation machine," that is, following Dahler-Larsen's definition (2015), a structure or a function without any subjective or human representation that "lives its own life."

The task that I have set for myself in this book is to show how the evaluation of both the political institutions of the state and the knowledge produced by researchers working in them became an inevitable part of the research process itself. In addition, in this book, I draw attention to the consequences of these processes for academic labor. In defining universities (and at a broader level, all research-oriented institutions), I follow Pusser and Marginson (2013). They understand universities as political institutions of the state because they require state resources, the state provides them with certain benefits, and universities in turn gain some authorities from the state. Both public and private institutions can be understood as political institutions of the state because many governments position universities as part of the state's portfolio of responsibilities. However, given that research, innovation, and knowledge are crucial for the development of states, this relation is complex

and goes beyond the mere provision of benefits and authority. From an economic point of view, the science sector is thus a strategic one. According to the Triple Helix thesis that describes relations between university, industry, and government, universities can play an enhanced role in innovation within increasingly knowledge-based societies (Etzkowitz & Leydesdorff, 2000).

Research evaluation systems are science policy instruments that consist of the sets of protocols, measures, indicators, and policy aims that are used for assessing the research productivity and activity of political institutions of the state. Research evaluation systems are some of the key instruments used in various countries, including Australia, Argentina, the Czech Republic, Finland, Norway, Italy, Poland, and the UK. In other instances, they are perceived as performance-based research funding systems or their key constituents.

In the ongoing discussion on research evaluation systems, the systems in these different countries are named as homogenous, examples of the same type of instruments (Hicks, 2012; Zacharewicz et al., 2019). For instance, the Research Excellence Framework (REF) of which the first forerunner was established in 1986 in the UK is indicated as the first research evaluation system, and then other systems are enumerated chronologically. These include, for instance, the Comprehensive Evaluation of Scientific Units in Poland, launched in 1991, and the Research Quality Framework (now replaced by the Excellence in Research for Australia) launched in 2005. Moreover, the advent of the first research evaluation systems is traced back to the 1980s, and their roots are connected to the spread of the tenets of New Public Management, global competitiveness in science, and the knowledge economy (Hicks, 2012). Thus, in the early policy statements presented by governments implementing research evaluation systems, one can find mention of numerous similar themes, including the distribution of state funding, the internationalization of research, and the general pursuit of excellence.

In this book, I want to critically consider the one which is taken for granted in studies on research evaluation and national science policies, that is, the homogeneity of research evaluation systems in terms of the conditions of their formation. While one can identify similar rationales presented in various countries as arguments for establishing research evaluation systems (e.g., funding distribution and the improvement of research productivity), this does not justify putting those different systems into a single category.

It is useful, for the purposes of this investigation, to look into some of the similarities between systems in terms of the protocols, measures and indicators used, and the policy aims. On the one hand, the similarities allow us to compare systems and investigate how they are constructed, including, for instance, what kind of publication counting methods they use. At the same time, such a comparison even allows us to analyze how these systems influence the productivity of universities in terms

of the number of publications. On the other hand, however, at a superficial level, the similarities conceal actual and deep disparities that are the products of the different contexts in which research evaluation systems were elaborated. Understanding and uncovering these differences enable us to show that the current research evaluation systems vary not only because they use different metrics and set different policy goals, more importantly, in light of this book's aims, they differ because some of them were established in the countries of the Eastern Bloc, where the so-called research evaluation systems had existed before the advent of New Public Management.

At first glance, any two research evaluation systems can appear similar. For instance, the Australian and Polish systems use similar journal rankings and discipline classifications, and today even societal impact is assessed in a similar way in these two countries. However, the context and the conditions in which these systems were established are different.

The Polish research evaluation system was established in 1991 just after the start of Poland's economic and social transformation from what was later termed "real socialism" to democratic society and the free market. Therefore, the Polish system was one of the key policy instruments that served the depoliticization of research and the implementation of objective measures within the higher education and science sectors. Given the goals of the process of transformation, establishing a new research evaluation system was imperative as a condition for moving away from a centrally planned economy. My use of the word "new" for describing the system is intentional. In Eastern Bloc countries, not only were science and the economy centrally planned, so too was the research evaluation system which served to promote the realization of socialist science goals. And yet, despite these singular characteristics, this chapter in the history of research evaluation and the measurement of science is all too frequently omitted.

The Australian research evaluation system was launched as part of a five-year innovation plan called Backing Australia's Ability. This overall state strategy aimed to build a knowledge-based economy and to enhance the government's ability to manage the higher education and science sectors in the global context. The goal of assessing research quality and the impact of research was to provide an answer as to whether public funds were being invested in research that would deliver actual impacts and provide benefits to society. In this perspective, it is beyond question that New Public Management is crucial for understanding the background against which the forerunner to the Excellence in Research for Australia was designed. Key concerns at the time were public accountability for resources and the search for the most effective way of determining the allocation of funding, at the time of budgetary restrictions, to Australian universities. The context in which the Australian system was designed is similar to that in which the first and later versions of research evaluation systems were established in the

UK. Tracing this move from accountability for public funds to the assessment of research excellence and the impact of research, one can discern the policy aims that have been prioritized by the state. Still, accountability for public funds and ensuring that they are invested in research that can benefit the wider community continue to be key elements of the environments in which research evaluation systems are designed in both Australia and the UK.

These two examples of research evaluation systems highlight the contrasting climates in which the systems, measures, and metrics were drawn up in Australia (and the UK) and in Poland. More significantly, these diverging contexts have had an important impact on how the same indicators or methods of evaluation come to be perceived differently by the academic community in these countries.

For instance, peer review is usually invoked as the best way of evaluating research results and the impact of research. Peers are treated as the key pillar of science, even though criticisms of peer review are occasionally raised. Thus, the peer review practices implemented within the research evaluation systems in Australia and the UK are presented as benchmarks for other research evaluation systems. In relation to this, the well-worn argument is made: While metrics might be useful, only peers can actually evaluate research and its impact. And yet the key factor in peer review, that is peers themselves, can also be perceived as the greatest weakness of the research evaluation systems, which metrics can be understood as counter-balancing.

In the post-socialist countries of the Eastern Bloc, peers were the hallmark of centrally controlled science: Their decisions were based not on merit but on the political agenda. Wouters (1999) cites an extract from an interview with A. A. Korennoy – a PhD student from Gennady Dobrov and one of the founders of scientometrics in the Soviet Union – who explains that even scientometrics' analyses of research efficiency and productivity were not used to inform policy decisions and funding: "The decisions taken were mostly voluntaristic and guided by completely different considerations. The funds were allocated not according to the front of research but according to personal acquaintanceship" (Wouters, 1999, p. 92). Therefore, one of the ways of making central planning in science a thing of the past during the transformation period of the 1990s was to rely on metrics that were perceived as objective that is not dependent on peers' decisions. Thus, for example in Poland, researchers did not trust their peers and preferred metrics (Mishler & Rose, 1997). Accordingly, each new version of the Polish system was more and more metric oriented. However, neither Polish researchers nor scholars investigating research evaluation systems seem to have noticed that the process of constructing metrics is one in which peers and political agendas are very strongly involved and that ultimately metrics are a hallmark of state power.

Resistance against metrics used in research evaluation systems also differs across countries. In the leading countries for research like Australia, resistance

against metric systems takes a different form from that in peripheral countries (Beigel, 2021; Kulczycki, Rozkosz, & Drabek, 2019; Woelert & McKenzie, 2018) because metrics – most often designed by central countries – confirm and legitimize the leading position of those countries. In other words, when metrics are based on data that favor publications in English, it is unsurprising that resistance in Australia or the UK should assume a different nature from resistance in countries like France, Italy, or Ukraine. On this point, I am not claiming that Australian researchers do not resist the use of metrics (Hammarfelt & Haddow, 2018). Rather, I argue that their reaction is shaped not only by science systems and metrics themselves but also by the cultural and historical context.

In this book, I argue that it is not only the case that different sets of metrics, in their implementation, lead to different consequences. It is equally important to consider the process through which metrics are constructed because this process is always situated in a specific context, place, and time. Thus, although two different countries might use the same metrics, they may have been constructed for completely different reasons. Accordingly, in this book, I explore how the state constructs measures and indicators and imposes their use on universities and research institutes.

A research evaluation system is a technology of state power which serves to sustain a power relation built by, on the one side, the state and its agencies and, on the other side, universities and researchers. This technology of power in science transforms the production and communication of scientific knowledge; thus, research practices are influenced by various metrics used in research evaluation systems. For example, using an impact factor for scientific journals can simultaneously encourage researchers to publish in top-tier journals or prioritize quantity over quality in relation to their publications. Investigating the actual (un)intended effects of research evaluation systems is a complex task in which many factors need to be taken into consideration (e.g., gross domestic expenditure on R&D, number of researchers, and policy aims). Existing analyses have revealed many interesting dependencies across funding levels, the metrics used, and the policy aims prioritized. Despite this, many questions related to the rise, development, and role of research evaluation systems remain to be addressed.

1.2 The Evaluative Power of the State

Even authoritarian governments seek international legitimization of their actions. This is why Frédéric Joliot-Curie and Irène Joliot-Curie were invited to participate in, and thereby sanction the event at which the state transformed the academic landscape in Poland. The state had various political and policy tools at its disposal with which to achieve this, all of which were manifestations of its power.

The state has always distributed resources and regulated the public sector. However, the way in which this is organized has changed significantly over the past decades. When, at the beginning of twentieth century, Weber (1978) described how the rationalization of society produced bureaucracy, the characteristics of the public sector were distinct from those of other types of organization. It was only in the bureaucracy that roles were separated from persons, structure organized in hierarchical manner, favoritism eliminated, and a regular execution of assigned tasks implemented. A century later, in many countries, public administration is organized in a manner similar to corporate or private institutions that use key performance indicators, contracts, and a linear model of input–output budgeting, while implementing accountability systems around resource use (Dunleavy & Hood, 1994). This is because, as DiMaggio and Powell (1983) argue, the rationale for bureaucratization and rationalization have changed. Today, the state has become the evaluative state (Dill, 2014; Neave, 2012). The reform of public administration, from bureaucracy to the evaluative state, has been identified with the rise of so-called managerialism or New Public Management (Hood, 1991), which describes a particular way in which relations across government, public institutions, and society have been transformed.

Through varied historical transformations, universities have been confronted with diverse new expectations about their missions, tasks, and organization. The classic conception of the university is most often connected with the idea of the Humboldt University whose structure was defined by a set of autonomous chairs with students affiliated to them. Such universities were autonomous in the sense that professors (chairholders) were autonomous in terms of teaching and research. Since then, the idea of the university has changed many times, and universities have been understood, among other things, as public agencies, corporate enterprises, or innovation-oriented institutions. Autonomy within the university has also been redefined and today one encounters two main approaches: The first one still promotes the autonomy of academic staff members as in the Humboldt university, whereas the second highlights the autonomy of university leaders to define and realize university strategies and to manage the institutions. These approaches emphasize the mission (teaching vs. research) or the autonomy of the university (academics vs. managers). Nonetheless, they do not frame these institutions as implicated within power relations that derive from their dependence on the public funds distributed by the state. And yet today universities and other political institutions of higher education, together with the science sector, are influenced by science policies, including national ones.

As a concept, the evaluative power of the state can be used in many forms and contexts. Using it productively requires that one both identify and prioritize key aspects of power. Even then, however, given that the concept of power is a very complex

one, the likelihood remains that one is charged with being imprecise or unclear. In this book, my focus is on investigating how the power of the state transforms the production of scientific knowledge and the very research practices themselves. As I argue in Chapters 2 and 3, this power impacts on political institutions and researchers both directly and indirectly. In terms of its direct impact, this is exerted through state regulations, policy documents, and policy decisions. Its indirect impacts are realized through the shaping of the conditions and environments in which researchers work, which include labor conditions, types of employment and contracts, methods of assessment of academic staff, and the amount of financing distributed to universities.

The key causes of these indirect impact are the technologies of direct impact, that is the above-mentioned regulations, documents, and policy decisions. Nonetheless, both kinds of impact produce both intended and unintended effects simultaneously. In other words, the technologies of direct impact can influence the productivity of some researchers, bringing them into line with policy aims (e.g., the increase in publications in top-tier journals), while at the same time, another group of researchers can transform their publishing practices in unintended – from the science policy point of view – ways (e.g., more publications in local scholarly publication channels). Additionally, the shaping of work environments as the effect of indirect impacts can improve researchers' productivity and their focus on the societal impact of research which might be an intended effect of policy regulations. However, the indirect impact of, for instance, exactly reproducing national evaluation procedures at the university or faculty level can lead to a deterioration in the quality of academia as a workplace and, as a consequence, reduce the innovativeness of research. Such an effect would be unintended from the science policy point of view.

In order to investigate the effects of state power, we must go beyond the assumption that state power is a very complex mechanism embodied in power relations between the state, political institutions, and the researchers working within them. We must also view it as a set of actions and state capacity as mediated and implemented by numerous technologies of power or policy regimes. Hence, in examining the power of the evaluative state, it is necessary to combine two – at first glance – antithetical perspectives: Foucault (1995) and Lukes (1974)' definitions of power. These two conceptions can be perceived as antithetical because while for Foucault, power is an unintentional and overarching condition, for Lukes, power is a person's capacity to influence or change the interests of someone else. Lukes' definition of power as a capacity to do something highlights its intentional dimension. Whichever definition one adheres to, the following question needs to be addressed: Can the power of the state be both unintentional and at the same time, intentional? If one defines power as power relations and technologies and then focuses on the effects of the use of these technologies, one finds that in order to understand power itself, one must understand the (un)intended effects produced

by power technologies. One should therefore conceive of power as both an overarching mechanism and as the capacity to change someone's interest and action; doing so enables us to investigate power effects in a holistic way.

Foucault (1995) argues that power designates the complex and all-encompassing condition that produces and shapes our social reality, actors, objects, and relations. Understood in this way, power is not intentional action or strategy: It is because of its complex nature that power can change us, and not because some actor or institution that "has power" decided to do so. Thus power is not something that actors can have but is instead a complex relation in which they are involved. People and institutions involved on a continual basis in such situations internalize external control that makes them more or less willing to subject themselves to the societal norms and expectations that are the product of power relations.

Through its varied technologies, power colonizes people's minds. It is disciplinary power that becomes embedded in the various administrative routines of institutions. As presented in *Discipline & Punish*, disciplinary power is connected with closed spaces like prisons, hospitals, or schools in which control was exerted together with the restriction of freedom. In the era of New Public Management, however, these institutions have been transformed. Therefore, the logic of power itself has also been changing: It no longer functions under this disciplinary modus operandi but rather relies on the semblance of freedom coupled with uninterrupted control. In the books, he wrote after *Discipline & Punish*, Foucault argued that it is not possible to study the technologies of power without also considering the political rationality that underlies them. Thus he coined the concept of "governmentality," which combines the perspective of the state that governs others and the perspective of the selfthat governs itself (cf. Lemke, 2002). In this optic, subjects treat external norms as their own and govern themselves in order to meet external expectations. This observation is critical in my consideration of scholars' attitudes to various science policy instruments and their resistance to power structures, which are presented in the next part of this book. Although power's effects cannot be resisted, subjects are nonetheless aware of power relations and technologies and can, at least hypothetically, try to resist power.

Power, according to Foucault, is an unwilled complex mechanism that affects individuals and institutions. Steven Lukes, with his radical view of power (1974), defines it in a different way. Lukes shows that in the past, power was mostly understood as a one- or two-dimensional capacity, and his argument is that it should instead be defined as three-dimensional capacity. The one-dimensional view of power defines relations between persons or institutions as the capacity to convince a person to do something which they would normally not do. The two-dimensional view redefines relations and emphasizes the idea that having power means having the capacity to put up obstacles and, in this way, reduce and control others' options. As a way of developing these two approaches, Lukes suggests that one

characterize three-dimensional power as a person's capacity to influence, shape, or change another person's core interests.

In this view, then, power manifests itself through domination, that is acts of influence and manipulation. A person who influences or further, alters the interests of another person is an influencer who is working to promote an agenda. To put it differently: Power through domination is an intentional action taken by a specific person, by people, or institutions. While an influencer might not be recognized as such by those who are being influenced, it is still possible to reveal the power relations between them. In this way, Lukes' (1974) approach, contrary to Foucault's, highlights the intentional dimensions of power and focuses on decision-making and control over the political agenda (p. 25).

As a complex apparatus, power requires structures that need to be sustained. In societies, power's close relationship to knowledge is key to its reproduction. Foucault defined this relation through the concept of Power/Knowledge, where knowledge and information refer to individuals, groups, and institutions. It is the collection, archiving, and analysis of such knowledge that allows power to sustain and reproduce itself. As Weber (1978) showed, collecting and archiving information is one of the key characteristics of bureaucracy. In this way, bureaucracy and – for the past decades – evaluative states use collected and archived knowledge to control and govern, among others, the public sector. Thus knowledge and information become power technologies of the state.

Building on Lukes' conception of power, I define the evaluative power of the state as the capacity to influence or transform the interests of individuals or institutions and to modify their practices and behaviors. Evaluative power is reproduced by the very context that it itself produces. Thus, its capacity is realized not only through the implementation of policy instruments but also through the redefinition of the context. The concept of the evaluative state as described above underlines the fact that the state produces power relations by implementing various evaluation instruments. However, in this conception, the emphasis is mostly on the intentional actions that influence (state administration, policy makers). It is my contention that by bringing together Foucault and Lukes' approaches, we can deepen our investigations of the (un)intended effects of research evaluation systems.

From Foucault's perspective, disciplinary power enables the sustenance and reproduction of all-encompassing power relations; however, following the logic of this model, it is not possible to identify actual agents or influencers, as disciplinary power influences and transforms individuals and institutions themselves. One can therefore conceive of the inevitable and inescapable evaluation of science by and, as a consequence, governance of science by the evaluative state as a form of discipline in Foucault's sense. Evaluation, like discipline, is based on normalization and constant surveillance and drives both individuals and institutions to continuous

self-evaluation, and to comparisons between themselves and other self-evaluating entities that are also subjected to this all-encompassing mechanism.

In my investigation of the evaluative power of the state, I build on the idea that the intentional dimension of the power (in Lukes' sense) of the evaluative state needs to be examined in combination with power's unintentional dimension (in Foucault's sense), in which evaluation is understood as a form of discipline. In this book, my core concern is with the concept of evaluative power rather than with that of the evaluative state. This is because I am interested in the following two areas: (1) how the state transforms scholarly communication by constituting evaluative objects and by evaluating political institutions and (2) how scholarly communication is transformed by various self-evaluation practices (resulting from the internalization of evaluation norms) and forms of reactions and resistance against evaluation itself. While the concept of the evaluative state highlights the capacity of the agent, that is, the state, to manage and govern through diverse evaluation regimes, the concept of evaluative power focuses on power relations (and not the agent's capacity) that are cocreated and mediated by the agent and its technologies of power.

If a person is in the position to decide whether to measure, this implies that they hold power. This power might also be strengthened if they can also determine the way in which that measuring occurs. In publication-oriented academia, the measure is well known: it is the publication itself as characterized by various numbers such as the number of citations or social mentions or the opinions of peers and experts. However, as I explain below, measuring is not the measure, and measuring is linked with deciding how this measure is used depending on "what" and "who" is measured.

Witold Kula (1986), in his *Measures and Men*, reconstructed the social processes involved in constituting varied measures and ways of measuring. In feudal society, there was the widespread view that it was legitimate for a tradesman to use one measure when buying and another when selling. However, even one measure could have two different types of use. For instance, when a merchant was selling a bushel of grain, the bushel was struck (strickled) or the grain would be leveled with the bushel's rim. Yet when someone repaid the grain to the same merchant, the bushel needed to be heaped or "topped up," simply because the merchant had power to enforce this (Kula, 1986, p. 103). Here then is an example of the use of a single measure (the bushel) in which the quantity of grain differs because the power relations are different. Thus one can say that power manifests itself in the imposition of a method of measurement.

Evaluative power in science is embodied, among other things, in research evaluation systems, funding agencies, and varied accreditation procedures. Where the state has public control over institutions, there always exists some form of evaluation which is inevitable and inescapable because of the very nature of the evaluative state (Neave, 1998). In the science sector, evaluative power designates the power

relations produced by the state across political institutions, researchers, and state officials and policy makers. These power relations manifest mainly in (1) the design of measures, (2) the use of these measures to evaluate political institutions and researchers and to make a range of decisions using the evaluation results, and (3) the reactions and resistance of researchers against evaluative power and its effects.

These three manifestations of power relations determine the three main lines of inquiry pursued in this book: national science policies, research evaluation systems, and the evaluation game. The intertwining of these areas produces tensions across all parties implicated in power relations: the state, academia, and researchers. These tensions are the basis for resistance against the imposition and use of measures to evaluate scientific work. Finally, these tensions produce the evaluation game in which the rules and stakes revolve around measures and measurement

1.3 Games as Redefined Practices

In order to investigate the practices of any group or community, we must determine how certain actions can be identified as actions of the same type. Doing so allows us to pinpoint why specific activities carried out by different people should be perceived as actions sharing a common denominator. In other words, to specify why these actions constitute a social practice and how the meaning of this practice and action is reproduced in society.

Science is a cultural practice and as such it consists of rules, norms, values, standards and, in a more general sense, knowledge that are shared by members of a given community. Thus actions can be understood as realizations of a given practice (e.g., writing this book as a practice of scholarly communication) when a person follows specific rules (values, standards etc.) shared by a community of scholars (e.g., a manuscript should present the research results, and relevant works from the field should be cited). Action alone or even many actions alone do not constitute a practice because every practice is oriented toward interpretation. For example, this means that the specific action of writing this book is a realization of a practice of scholarly communication (i.e., its meaning is determined by a given practice) only when other researchers can interpret my action (writing a book) and its results (this book) in light of the rules and values shared by researchers. Practice is meaningful when either its actions or its results are communicated and accessible to other members of the community. For instance, if novel and clear argumentation in scholarly work are important values for researchers, then researchers who read this book can interpret my writing and assess whether the rules, norms, and standards have been properly followed in the attempt to realize these values. However, an unpublished book is an output of writing action but not of the practice of scholarly communication because members of the academic community cannot assess and interpret it.

Researchers are always involved in various scholarly practices at the same time. They do research, write papers, analyze data, evaluate proposals, manage institutions, organize conferences, communicate with peers, and realize countless other practices to which they are socialized by taking actions and by experiencing their effects through feedback. The meaning of their practices and actions is grounded in values that determine what the best path is for realizing a given value. In other words, what norms one should follow and according to what rules one should act. In practicing science, however, researchers often have to assume a dual identity or dual loyalty because of conflicts between the values that drive their actions.

Researchers have to realize numerous practices that are specific to the institutions in which they work or to the scientific discipline to which they belong. Loyalty to the institution in which they work is always a local form of loyalty, but loyalty to their discipline is always global because, by its very nature, science is international. Therefore, a point of reference for assessing the value and meaning of someone's actions (e.g., writing and publishing a paper) can be set either locally or globally. This implies that some actions can be in line with the values shared by researchers employed in a given institution and simultaneously, out of step with the values shared by researchers in a given discipline. For instance, because research is international, within many disciplines the best (or even the only) way to communicate research results is by publishing them as a journal article in a top-tier journal. This is a practice grounded in the value of promoting the broadest possible communication with peers around the globe. This value of global communication is set for all researchers and all actions taken by them, that is their publishing activity, are interpreted in light of this value. Nonetheless, from the second half of the twentieth century in various European countries and in the United States, one can encounter the practice of publishing a *Festschrift.* Let us take a look at this practice to see how conflict between loyalties can occur.

A *Festschrift* is a scholarly book (most often an edited volume) honoring a respected scholar and published during his or her lifetime. Editing a *Festschrift* or contributing to it (by writing a book chapter) is a way in which colleagues, former students and friends can pay homage a researcher. Most of the time, a *Festschrift* consists of original contributions prepared especially for the book, although occasionally, contributors may submit papers that are difficult to publish elsewhere. The practice of contributing to a *Festschrift*, which is often published by a publisher with local distribution only, produces a tension between the global and local loyalties of researchers. On the one hand, to write a book chapter is to go against the standard practice of the best way of communicating research in one's discipline. On the other hand, writing a book chapter for a *Festschrift* is an appropriate way of cultivating values shared by colleagues from an institution. In the course of everyday academic work, all researchers face such dilemmas and tensions,

because their working conditions are shaped mostly by the local context in which their institution operates. And yet recognition of their work by their disciplinary community is grounded not in their local, but rather in global terms.

Given that their actions are driven not only by a desire for recognition but also by the need for stable and healthy working conditions, researchers act and practice under parallel (and sometimes mutually exclusive) value systems that produce multiple tensions between their local and global identities. Such tension between two loyalties can also be understood in the light of the concept of the evaluation gap described by Wouters (2017), that is, as tension between what researchers value in academic work and how they are assessed in formal evaluation exercises. There is no indicator, as Dahler-Larsen (2022) argues, that could finally close this gap. However, the tensions between the values of an academic community and those grounding evaluation systems are not the only ones that affect researchers working in academia.

Academia is further subject to tensions generated by power relations across the global and national planes, and between institutions, decision makers, and researchers. When a state produces national regulations for research evaluation, all of academia in that country needs to situate itself within this new environment that has been produced by evaluative power. In sum, while the state evaluates, academia is evaluated and, on being evaluated, researchers react.

It is not, however the case that academia only follows state regulations (e.g., collects and archives information, calculates statistics, and assesses researchers). It also reacts to evaluative (disciplinary) power through various forms of adaptation, resistance, and struggle. As Foucault argues: "people criticize instances of power which are the closest to them, those which exercise their action on individuals" (Foucault, 1982, p. 780). In the case of research evaluation systems, this manifests itself in the fact that people in academia focus their criticism on the state administration and burdensome nature of reporting about their work, rather than the basic conditions that allow evaluative power to arise. Among such conditions, I include the rationalization of society that fed into New Public Management, academic capitalism, audit culture, and the neoliberal university. These necessary conditions for the existence of the evaluative state produce an all-encompassing mechanisms of power. Thus, while they are invisible for those in academia during their regular work, discussions, and reflections on academia itself, their consequences are nonetheless felt at work.

It is only through critical investigation aimed at uncovering power relations that one can elucidate the fact that any effort to change academia needs to take into consideration not only current technologies of power but also the conditions that make those technologies possible. While revealing power relations is one of social research's critical tasks, it should not be its ultimate objective. Investigators focused on research evaluation should go further and recommend next steps that can deepen our understanding and improve the situation of all (or at least, some

of) the parties implicated in power relations. Doing this, however, requires a better understating of the reactions, responses, and resistance that surface within academia when it operates under evaluative power.

In this book, I argue that resistance in academia against evaluative power manifests itself through various forms of the evaluation game. Evaluative power is reactive because it causes those working in academia to think, act, and react differently (cf. Espeland & Stevens, 2008). I use the term "game" because evaluative (disciplinary) power produces an all-encompassing situation that is nevertheless rule-based and has defined the ends. Moreover, I build from Foucault's idea that relationships of power are "strategic games" between liberties in which some people try to determine the conduct of others (cf. Lemke, 2002).

In defining the game, I construct a framework in which players (e.g., political institutions, researchers) are socialized for the game by taking actions (e.g., writing manuscripts, planning research) and by experiencing their effects through feedback which is deliberately built into and around the game (cf. Mayer, 2009). In terms of the rules of the game, these are set by those who have power in the power relation. The idea of game is one of ways of conceptualizing social interactions in any environment in which rules are explicitly stated and in which rule-makers might, to a certain extent, be identified. This approach has a long tradition in social sciences and one can point to many different perspectives, among them, George H. Mead's (1934) interactionist approach, Thomas S. Szasz's (1974) perspective on games as a model of behavior, and Erving Goffman's (1972) approach in which the game is the context in which behavior takes place.

Within studies of academia, "playing the game" is not a pejorative term akin to the idea of "gaming" but rather a name for the day-to-day practices of academic labor within a rule-based environment. Bourdieu uses the concept of game to explain the meaning of the field and argues that a game has no explicit or codified rules (regularities) but is worth playing (Bourdieu & Wacquant, 1992). In writing about evaluating the evaluation game, Elzinga uses the term *evaluation game* to analyze a methodology of project evaluation and writing of the possible effects of evaluation on research practices. Kalfa et al. (2018) using Bourdieu's understanding of game explore how academics respond to managerialist imperatives within the academic game. Lucas (2006) describes research in the competitive global market as the *international research game*. Fochler and De Rijcke (2017) use the term *indicator game* to name the "ways to engage with the dynamics of evaluation, measurement and competition in contemporary academia" (p. 22). Blasi et al. (2018) argue that evaluation can be treated like a game played by the institution against the actor: "the actors receive a payoff from their behavior according to predefined rules and will engage in the strategic games to beat the rules. The extent to which strategic games can be played depends on the nature of the rules" (p. 377).

Yudkevich et al. (2016) use the concept of the game as a framework for their edited book *The Global Academic Rankings Game* and define the *rankings game* as a high-stakes exercise that exerts influence on institutions' internal policies.

Suits (1967) demonstrates how games are goal-directed activities in which inefficient means are rationally chosen. For instance, while playing soccer, no one can touch the ball with their hands unless they are the goal-keeper. When a player is kicking the ball toward the goal and another player (not the goal-keeper) decides to use a leg instead of a hand to block the shot (although the latter would be much more effective), then he or she is rationally using an inefficient means which is permitted by the rules. Thus, in all games, we need to know what means we can use and what means would be classified as rule-breaking.

How do games – which are also social practices – differ from other rule-based practices? Suits (1967) argues that people obey the rules simply because such obedience is a necessary condition to make *playing the game* possible. In other words, following the rules makes it possible for the game to take place and thus we follow the rules for this purpose. In other types of practices and activities, there is always another (external to the game itself) reason for conforming the rules. If we consider, for example, moral actions, following the rules (e.g., rules derived from religious values) makes our action right and not following the rules makes our action wrong. An analogues situation can be found in communication acts. If, for example, the aim of the act is to respect, through mourning, those who recently died, people can follow the rules and engage in a moment of silence, which then makes the silence a gesture of respect. Not following this rule when others are makes our action (e.g., speaking loudly) wrong in the context of this specific communication practice.

A social practice – such as language – cannot be created in a short period of time by decision or by the act of one or a few persons. For instance, researchers write and publish journal articles because in their disciplines or institutions this is how science has been done for years. It requires a great deal of work and time to change such a practice through a bottom-up approach, which is to say by common decision of the practitioners themselves (e.g., researchers from a given institution who practice research). This would apply, for instance, to a desire to shift the practice in order to start publishing more internationally oriented papers or to publish in English when an institution is based in a non-English–speaking country.

One cannot reduce the process of altering a practice to the presentation of new aims and goals (e.g., "our institution needs more publications in English") and implementing new techniques (e.g., "researcher can attend English lessons"). Such a process requires a change of mentality or rather of the collective representations shared by a community and embedded in their practices. When an institution or discipline decides to modify a practice, it initiates a process that consumes a great deal of time and work.

A change (even a radical one) of practice or the implementation of a new practice can also occur fairly rapidly through top-down approaches, that is through the implementation of a new set of rules or by forcing a change in the current situation. For instance, this can occur when a state which, for several years has been evaluating institutions according to the number of peer-reviewed publications they produce, informs institutions that from now on, they will be evaluated and financed according exclusively to the number of peer-reviewed publications they produce in English. In such a case, one can say that the norms, rules, and values of academia have been subjected to rapid transformation. From the perspective of researchers and institutions, rule makers introduce new regulations as if they were pulling instructions from a new boardgame box: from now on you have to follow these rules because you work in a political institution of the state. In such cases, the implementation of new rules by evaluative power will appear abrupt, even if there has been a process of public consultation, or of preparing the new regulations. The outcome of such processes is that a completely new situation is produced.

Finding themselves in a new situation, researchers and managers have to gage how to act in order to comply with the criteria and values defined by the new rules. They do not only look for the most efficient way of practicing science or communicating their research but also start to think about what means they need to use so as to secure their position in the new (evaluative) situation and to follow the new norms. If in a new regime of rules, only English language peer-reviewed publications count, then researchers, for example, those whose main disciplinary language of publication is German, will start to evaluate how to go on with their work and publishing. In this instance, according to the values shared within their discipline (discipline loyalty), the best way to publish is to publish in German. And yet according to the values shared by managers from their institution, it is best to publish in English or in both English and German. In this way, that is by a rapid implementation of new value regimes, multiple tensions between researchers' loyalties are created. If I am a researcher who needs to decide how to act in such a situation, I can choose one of the following strategies: (1) I am loyal only to a discipline and publish only in German; (2) I am loyal only to an institution, and I stop publishing in German and start publishing in English; (3) I try to be loyal both to a discipline and to an institution, and I publish both in English and German, even though publishing in German is perceived as a waste of time and resources.

Researchers and managers are regularly confronted with such decisions. In this way, academic work becomes a game, that is, a situation that is produced by a top-down implementation of new rules related to a social practice that has been cultivated for many years and has established norms, values, and rules. Introducing new or modified rules initiates the process of institutionalizing the practice and establishing the game. When a game coheres, researchers and managers start to

play; each of them has their own individual strategy, yet at the same time, interactions between players can also modify the way in which they play.

A change of rules provokes researchers and managers to start searching for ways of engaging in practice (in Suits' sense), which would allow them to follow the new rules at the lowest possible cost, because the rules can be changed again at any moment. This redefined practice is realized by bringing actions into line with the new rules, which does not necessarily mean that they are in line with the values and aims of the institutions, disciplines, or even the state and rule-makers. What matters is that one meets the criteria laid out by the new rules, whatever the cost. And the only reason for conforming to the rules is in order to reproduce the situation in which one can follow the rules, because reproducing them is what allows one to be employed in a given institution. This is why I call the *game* a specific form of social practice: It results from a rapid top-down redefinition of ongoing practices by the implementation of new (or modified) rules through various technologies of power (e.g., research evaluation systems).

The game is not a typical bottom-up social practice; it is rather a top-down redefined social practice. The power to rapidly redefine a social practice is always external to practitioners. Practitioners become players because they want to at least maintain their current position. They play because there are resources at stake in the game (e.g., stable work conditions, funds for research), with the goal of game being to win and gain resources. As long as an evaluation system is operational, the evaluation game does not end. While the stakes of the game vary, depending on various conditions, with players having different starting points, the key issue is that once the game is established, those working in academia are forced to play it. Therefore, they try to adapt and modify their practices in line with the new rules and aims implemented by the evaluative state.

An adaptation constitutes a *strategy* as to how to play the game and, in this way, scholarly practices become evaluation-driven practices in academia. A strategy is not something that is intrinsic to a game but rather something that a player brings to the game. As Avedon (1981) argues: "it is something that the player develops, based on his past experience, knowledge of the game, and the personality of the other players" (p. 420). Thus, different strategies are used in different institutions, fields, or countries. People in academia can either adapt (or adjust) to the system, try to ignore it, or try to change it (cf. Bal, 2017).

1.4 Defining the Evaluation Game

Weigl's situation, described at the beginning of this book, can serve as an example of the evaluation game: He published a number of papers in order to hold on to his scholarship. This action was not in keeping with his loyalty to his discipline,

but he was compelled to take it by the all-encompassing situation that defined the conditions of his work.

The evaluation game is a practice of doing science and managing academia in a transformed context that is shaped by reactions to and resistance against evaluative power. Such a game is established in a dialectical process: Through evaluative power, the state introduces new rules and metrics, while researchers and managers in academia devise various strategies for following these rules at the lowest possible cost. These strategies – as forms of adaptation, response, and resistance against evaluative power – are reactions to new rules and metrics.

Playing the game does not interrupt social practices (e.g., communicating research results) but instead puts the accomplishment of the new goals and compliance with the new rules in first position. In this way, people may start looking for ways to get around these new rules according to which a practice is evaluated. If, for instance, an institution assesses an individual scholar based on the number of journal articles they published in a four-year period, then playing the game would involve an artificial increase in the total number of co-authored articles. Thus, two scholars who used to publish single-author articles could decide to start writing joint articles (or even to add each other as authors to papers actually written by only one of them). This would then occur not because they had started to collaborate, but in order for them to increase the number of articles to their names. Here the game – that is, reaction to the rules of evaluation – has just started. A practice of scholarly communication is still being cultivated (journal articles are published), but this practice takes the form of a game in which the end is not to communicate research results in the best way (from a disciplinary perspective) but rather to communicate the research results in line with the evaluation criteria and with the scholar and institution's interests.

The evaluation game manifests in the day-to-day work of all those involved in the diverse power relations of evaluation. Therefore, the game can be played by all parties in those relations, that is, (1) researchers, (2) managers, and (3) the policy makers who design the measures used within research evaluation systems. Let us consider an example to see how individual actors participate in the game.

Global competitiveness for key resources like funds, researchers, and students, as well as university rankings, exerts a pressure on governments and policy makers to improve the productivity and quality of research in terms of bibliometric indicators. As a rule, bibliometric indicators are only a proxy for research quality, and yet they are identified as the information that shows how well a given country, institution, or researcher is performing. Thus, such indicators play an important role in rankings, and in this way, global competitiveness contributes to the emergence of the evaluation game.

In this context, policy makers start to play. The aim of the game is to boost the position of universities in their country in rankings that are considered important

within that country. From the perspective of policy makers, the goal of the game is to legitimize the funding of the higher education and science sectors. It might be the case that policy makers decide on certain areas within the ranking criteria which they believe can be improved, and then focus only on these. For instance, in some rankings, only publications from the TOP 10% of top-tier journals in international databases like the WoS or Scopus are counted. Therefore, policy makers may respond by creating a research evaluation system for assessing and funding institutions. In such a system, publications in TOP 10% would have substantial weight in relation to other publications from outside of them, or indeed, other types of output. The policy makers' response is shaped by the global context of competitiveness and the expectations of the key stakeholder, that is, the government, that institutions from a given country should improve their position in the rankings. By designing an evaluation system, policy makers meet the expectations of ranking criteria. However, the effort focuses only on bibliometric indicators that do not capture the larger complexity of doing research. And yet in this scenario, policy makers have responded to a challenge at the lowest possible cost, that is they have focused only on what is counted and what might be improved in order to increase the position of institutions within a specific ranking.

In such a situation, how might managers of academic institutions react? They might play the evaluation game created by ranking pressures by applying the rules of institutional evaluation at the local level of researcher assessment. This practice is well documented in research evaluation studies and is called "local uses" (Aagaard, 2015). Managers may just copy the regulations, which is the move that entails the lowest possible cost. Thus they may permit researchers, as part of the individual researcher evaluation exercise, to report only those publications from TOP 10% journals.

In this example, the context in which researchers work has been redefined in a top-down manner. Researchers would know that in order to receive a positive evaluation in the upcoming assessment of their work, they would be expected to publish only in TOP 10% journals. In response, the strategies for engaging in the game would vary depending on the discipline in question. For instance, researchers from fields in which scholarly book publications play an important role might stop publishing books and try to redefine their research topics in a way which allowed them to publish in journals. Through such a strategy, publication patterns and publication channels might be transformed. In the disciplines in which publishing in journals is a common practice, researchers might decide to put more effort into publishing in better – from an evaluation game perspective – channels, or they might start to consider how to game the situation by publishing more co-authored publications that are not based on real cooperation.

There is naturally a great diversity of ways of playing the evaluation game. One might consider the consequences of choosing only TOP 10% journals in relation

to publishing in national languages, given that the majority of top-tier journals indexed in international databases publish only in English. I will present a detailed analysis of the different types and forms of games in Chapter 6. Here, however, let us summarize and pinpoint what this example tells us about the evaluation game as played by all parties implicated in this particular set of power relations.

As can be discerned from the above example, researchers play the game because a top-down redefinition of the context in which they work has occurred. When one analyzes how the evaluation game is played, one needs to take into account the fact that this game is constituted at all levels (global, national, and local) and that its actors represent all parties within power relations. Consequently, because the game is a social practice involving actors who follow (or not) the rules, and play for certain stakes, the game cannot be reduced or abstracted from the actors who play it, nor presented as existing at a separate level. Rather, one must view all these elements and parts of power relations as mutually constitutive. This is because the evaluation game is a response by all actors within the science system to power relations that are generated by the evaluative power of the state. This state power is, moreover, significantly influenced by the global context of doing and managing science; thus the higher education sector is shaped by global actors and institutions. In other words, one might say that when context is redefined through a top-down process, everybody plays the game. However, it should be noted that researchers, managers, and policy makers play the evaluation game while at the same time playing other games in academia that are related to teaching, their careers, and relations with peers.

1.5 Factors Contributing to the Evaluation Game

One can identify three main factors contributing to the emergence of the evaluation game in science. The first is related to the use of measures (quantitative indicators or metrics) to control, govern, or modify social behaviors in line with particular aims or targets that are external to those being controlled and governed. In such situations, an indicator becomes the target and in this way, the stakes of the game are changed. The second relates to changes in the context in which researchers and managers work due to the implementation of a research evaluation system or modifications to it. The third concerns tensions between the rationalities (or logics) adopted by designers of research evaluation systems and the rationalities actually used by people in academia. Let us look into these three factors in detail.

The first factor contributing to the emergence of the evaluation game is linked to the very nature of any measure deployed to monitor, control, govern, or modify social behaviors. Among other things, measures are used to provide an intersubjective assessment, and yet when they are used, they cause those working in academia

to think and act differently (cf. Espeland & Stevens, 2008). Donald T. Campbell made the observation – widely known as Campbell's Law – that any indicator used for social decision-making will become a poor indicator because of its very nature: "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor" (Campbell, 1979, p. 49). Similar research into the repercussions of using indicators, like Goodhart's Law, the Lucas critique or the Cobra effect, highlight what happens when numbers, quantification, and measuring are used to control behaviors and social actors: The indicator itself will become the target, and people will do what is being measured and stop doing that which is not being measured. In his *The Tyranny of Metrics*, Muller (2018) provides numerous examples of the consequences of using indicators in various sections of the public sector such as health, education, and higher education. All of these examples make it evident that if one decides to use numbers (indicators) to assess or monitor social practices, the practice itself will change. However, whether the change will be positive or negative for those subjected to these indicators is not something that can be determined in a top-down manner.

The second factor contributing to evaluation games in science relates to the influence of evaluative power on changes to power's all-encompassing mechanism and the day-to-day practices of academic labor. A change may consist, for example, in the introduction of a new way of reviewing publications submitted to a national evaluation exercise or in an increase in the number of publications that researchers need to present within specific time periods. In either situation, researchers and managers need to decide how to react. In an ideal world, managers would communicate changes in the regulations to researchers. Then, they would have a discussion on how to tackle the change, how it might influence the researcher's work, and what adaptations would need to be made in order to prepare for the change. In actuality, however, changes in evaluation regulations are often rolled out through long and drawn-out policy processes.

It is also the case that even the best academic managers are sometimes helpless in the face of vague policy documents and regulations: It is often difficult to understand how a minor change in regulations might actually influence the day-to-day practices of researchers or managers. Will it increase the burden of administrative work? Will the different scope of information to be collected in order to comply with the evaluation criteria require a new workflow? How might the implementation of a new bibliometric indicator (e.g., the Hirsch index) within national regulations impact on internal evaluation procedures at our university? Such questions follow each modification of the all-encompassing nature of power produced by evaluative power. When a change is made, then those working in academia have to operate in an unknown situation. Confronted with this situation, they start to think

about how to continue their day-to-day practices in the same manner as before and to change them only when it is unavoidable or absolutely necessary. Day-to-day practices in academia (defined in terms of "games") are usually long established, but a change in the rules can modify the stakes of the game (e.g., employment stability or funding for research). Moreover, a change in the rules can put the spotlight on the fact that that while those in academia all have to participate in the same game, they do so from different starting points, with some starting points, for instance, those of early career researchers, constituting a disadvantage.

The third factor leading to evaluation games in science consists in the tensions that exist between the different logics that motivate system designers and those working in academia. Those who design research evaluation systems have to make multiple decisions related to the following questions: What will be measured? (e.g., what kinds of activities will and will not be measured); who will do the measuring? (e.g., peers, stakeholders, and experts external to academia); who will be measured? (e.g., will outputs produced by PhD students be evaluated); how will things be measured? (e.g., should qualitative, quantitative, or both methods be used); what criteria will be used? (e.g., do only peer-reviewed publications matter, or are publications for the general public also important); when and how often should evaluations take place? (e.g., annually or once every four years); how will the information collected be used? (e.g., only for evaluation exercises or also for other administrative purposes); how will the results of the evaluation be presented? (e.g., through the ranking of institutions or only through information on positive/negative results); and finally, how will the results be used? (e.g., only for block grant distribution or perhaps also in order to change human resources policy).

While taking decisions on these questions, system designers necessarily adopt certain epistemological assumptions about the cognitive responses of those in academia who will be subject to the research evaluation system (cf. Pollitt, 2013). In other words, system designers assume that researchers and managers behave according to certain logics and in this way, they predict how researchers and managers will behave and modify their practices in response to the roll out of a research evaluation system. In his description of the logic of performance management regimes, Pollitt (2013) shows that two types of logic can be adopted by system designers: (1) a goals-oriented logic and (2) a logic of appropriateness.

According to the first logic, academics and managers are primarily motivated by wages and labor security. Thus, in the course of their daily work, they conduct research while at the same time trying to achieve the goals presented to them by stakeholders or managers. In this way, through the setting of targets, policy makers govern academia (cf. Bevan & Hood, 2007). Within the second logic, academics and managers follow the collective values and collective representations shared by their community. If policy makers want to govern academia, they need to not only

set targets but also remodel the environment in which academics and managers work. Such changes can then modify the set of values and collective representations shared by members of a given academic community. It is important to note that these two logics have substantially different consequences for the design of research evaluation systems. If designers assume the first logic as primary, then the system has to have clear goals, targets, and indicators that should be directly communicated to the evaluated community. In keeping with the second logic, academia also can be governed, but this requires more complex agenda-setting, time, and resources.

While these logics are adopted by policy makers and system designers at the macro level, system users, that is, researchers and managers, behave and practice at the micro level. Here, a much greater variety of logics are deployed, depending on the specific time and context (e.g., whether researchers work at a top research-intensive university or at a small local university). Pollitt terms such micro-level logics "alternative logics" (2013). Alternative logics can be adopted by single researchers, groups of researchers, or by the whole community at a university. Moreover, researchers or managers can follow diverse alternative logics at the same time. Pollitt argues that these alternative logics are used by actors who are subjected to performance management regimes. However, some alternative logics are in fact also adopted by policy makers and the designers of research evaluation systems. This is because they have to make assumptions about how academics and managers behave in their day-to-day work and are themselves also engaged within the wider system. That system involves the policies that are realized by the state, as part of which varied logics about people's behaviors are also adopted (e.g., how policy makers and policy system designers think or should think about social interventions).

Pollitt uses the term "alternative logic" to name the mostly unintended effects of performance management. Some of the examples of alternative logics he offers are the *threshold effect* (a minimum target can motivate those falling below the target but also de-motivate those who have already performed above the target), the *ratchet effect* (managers may be tempted to hit but not exceed the target if next year's targets are based on last year's performance), and *cheating* (not bending but breaking the rules). In other studies of performance management, one also finds diverse alternative logics that are also termed the "unintended consequences" of performance management (Smith, 1995) or of "gaming" (Bevan & Hood, 2007). In other words, alternative logics are responses to interventions and as such they vary depending on the specific context in which the interventions take place.

In Pollit's perspective, alternative logics are both logics at the micro level, that is, of people whose performance is being measured, and the consequences of the clash of macro and micro logics. I believe that distinguishing these two elements

of alternative logics, that is, the logic of a few people belonging to a group (micro level) and the consequences of the clash of macro and micro logics, enables us to achieve greater clarity when we lay out the second factor for the emergence of evaluation games. The consequences known as the *threshold effect, ratchet effect*, or *gaming* are not always the result of logics adopted by those who are subjected to performance management. Sometimes, the reason for which people start *gaming* is not because their alternative logic is different from that adopted by designers of the performance measurement system. It can be the case that these two logics are coherent but the situation in the workplace might change substantially thereby reframing power relations. Eventually, a new configuration of power relations might modify collective representations and the values shared by the community that is subject to the performance measurement.

In such cases, one of the above effects, like the *threshold* or *ratchet effects*, might be produced. However, these should not be conceived simply as micro logics pertaining to those in academia. They are, rather, diverse types of evaluation game, which involve not only players (those working in academia) but also specific rules and stakes. Moreover, the evaluation game often involves all parties implicated in power relations. For instance, the *threshold effect* is based on a rule about minimum targets: for example, the submission of three publications per researcher for every four-year period as part of the national evaluation exercise – as is the case in Poland (Korytkowski & Kulczycki, 2019). This target has the potential to motivate those researchers who produced only two publications in the previous period, but it could also demotivate those who produced nine publications. The latter might consider all their publications above the fourth as ineligible in terms of the evaluation exercise. However, this *threshold effect*, observed within a group of productive researchers, would not be produced only because of their adopted logics about "how to work" and "what good science is." The effect is better understood as an evaluation game caused by a redefinition of the context in which they work and the process of adaptation to it. Managers (i.e., a research organization in an institution) who use this particular regulation as the only stakes in the game are playing the game because they have the power to modify the stakes at their institutions by changing local stakes (i.e., increasing the number of publications which must be produced in a given period or introducing financial rewards for the best performers). Moreover, simply by virtue of the fact that they design and use the measures to govern and control, policy makers must also be viewed as involved in this game.

In further considering the three factors that give rise to evaluation games, one must answer two important questions. First, whether policy makers intend for these games to occur in academia and, second, which effects of research evaluation systems are intended as part of their implementation. In numerous studies on performance management systems, two terms, "effects" and "consequences," are used

interchangeably. Moreover, the effects and consequences of performance management are mostly presented as intended or unintended. Merton (1936) analyzed the unanticipated consequences of purposive actions and highlighted that unforeseen consequences should not necessarily be identified with undesirable consequences. The consequences of an action are limited to the elements in the resulting situation, which are exclusively the outcome of the action. When researchers' daily work is analyzed, it is difficult to separate out the specific factors that determine changes in the way in which they practice science, that is, due exclusively to factor 1, factor 2, or factor 3. The problem of causal imputation in comprehensive social practices and social interventions should be a warning to us: It is very difficult to argue that particular new rules have changed practice. One should rather say that it is the implementation of those rules that has influenced the practice or co-contributed to the changes observed.

Following Merton's argumentation, while the unintended and anticipated outcomes of actions may be relatively (as regards other possible alternatives) desirable for the actor, at the same time, they may be viewed as negative, in value terms, to observers or subjects of the action. This is to say that when the state implements a new research evaluation system, it will entail certain intended and anticipated outcomes and effects. Those effects might however be perceived by those in academia as negative effects of this implementation.

Let us look more closely at the *threshold effect* described above, in which researchers must submit their best four publications every four years for the purposes of the national evaluation exercise. An intended and anticipated consequence was to motivate those researchers who were slightly below the threshold. But can one say that the demotivation of those researchers who were above the threshold was an unintended and unanticipated consequence of the implementation of this rule? One can say that it was an unintended but anticipated consequence. System designers knew that some researchers might be demotivated, but the primary aim of the system was to motivate those (directly or through their institutions) who did not perform very well, rather than to keep motivated those who were performing well. Nonetheless, an evaluation game (i.e., the *threshold effect*) could occur in such a situation. Can one say then that the evaluation game was an unintended consequence of the implementation of new rules, or is it the case instead that it was a foreseen and anticipated transactional cost of the social intervention? This example shows how difficult it is to determine whether something is an unintended consequence of a specific action, especially where a person is trying to associate unintended consequences with unanticipated effects. In light of this, it is fruitful to consider Lewis' idea (2015) regarding the explicit and implicit purposes of performance measurement. When policy makers implement a new research evaluation system, they present various explicit goals (e.g., improving research performance,

motivating researchers to publishing in top-tier journals). An evaluation game, for example, the threshold effect, might be one of way of achieving the goals. Therefore, not all forms of evaluation game should be viewed as unintended and unforeseen consequences. Some might be intended or foreseen and in line with the explicit or implicit goals of the system makers. However, it is worth bearing in mind that the evaluation game is interactive among many actors, so one might justifiably ask why the intentions of policy makers should be privileged in assessing whether some effects are intended or no. I argue that their intentions are to some extend privileged because policy makers are real initiators and implementers of national or global research evaluation systems and as such play a key role in establishing the evaluation game. Moreover, drawing on the culturalist approach to science of Znaniecki (one of the founding father of science of science in the 1920s), the term "intention" does not mean here any mental state (of policy makers) but rather an action that is culturally meaningful and interpretable (Znaniecki, 1934). Hence, implementing a new evaluation regime can be understood as a communicative action, which involves policy makers (senders) and researchers and institutions (recipients). Thus, saying that some effect is intended means that some policy regulation has been received (implemented) and interpreted (influenced the practice of institutions or researchers) according to the goals of policy makers.

Dahler-Larsen (2014) argues that intentions themselves may not always be the best standard against which to assess the consequences of performance indicators. He therefore suggests that we use the concept of "constitutive effects" instead of unintended consequences, in order to be able to show that the use of indicators is truly political, because it defines categories that are collectively significant in a society. I agree with him that intentions are not the best standard and that differentiating effects based on whether they are intended or unintended is not only difficult to do but also does not provide us with useful tools for understanding what is actually going on with a particular social practice. Still, in discussions on the effects of research evaluation systems, policy makers and policy designers often claim that newly observed practices (which I would call types of evaluation game) were not intended. Hence, when they are under attack, policy makers often use the concept of unintended effects as a defensive argument. It follows that even if this distinction is not particularly useful from an analytical perspective, one should not totally give up on it precisely because it allows us to map policy makers' varied responses and reactions.

Merton (1936) drew attention to the fact that the consequences of an action do not only apply to those persons who are the target group of the action (and to the social structure, culture, civilization) but also to the actors themselves. In other words, the repercussions of a redefinition of the academic context extend both to academia and to those who use evaluative power to redefine it. Why then are actors

not able to foresee all the consequences and to be prepared for them? Merton argues that during actions, actors ignore facts or make errors in their appraisal of them because they hold certain interests that blind them to risk or which create self-fulfilling prophecies. These mistakes in turn generate unintended consequences that can lead to other problems – through a chain of consequences – and to new solutions connected with new, unintended consequences.

The research evaluation system spans an extremely diverse reality, for which reason it must prioritize certain areas at the expense of others. This complexity can also pose an obstacle to the translation of findings from performance-management studies (based on the private sector) to the academia. Even the largest company will not contain as many different tribes and territories as the academia does (cf. Becher & Trowler, 2001).

Some researchers may claim that they do not play the evaluation game in academia, and are simply doing good research. However, evaluative power transforms the context in which they work, which means that they need to play in order to maintain the previous situation. In other words, it is not possible to be employed by a political institution of the state and not play; both the general situation and evaluative power itself force us to participate. Can people choose *how* they play, at least to some degree? Yes, I would argue that it is possible to a certain extent to design a strategy for playing. However, one needs to remember that the game is conducted differently depending on the context; thus it varies from central countries (like the Netherlands or the United Kingdom) which have dominant and privileged positions in terms of resources of all kinds to other countries in a lower position in the global distribution of power (like Botswana or Puerto Rico), and again, to aspiring countries (like Poland or the Czech Republic). The stakes in the game therefore also differ. I will return to this question in a later chapter where I investigate the diversity of evaluative powers. I turn now to an examination of the conditions and context that lay the grounds for the rise of research evaluation systems.