From chasing violations to managing risks: origins, challenges and evolutions in regulatory inspections.

Blanc, F.O.M.

3. Theoretical underpinnings: costs and effectiveness, compliance drivers, discretion issues, risk and regulation

> *The starting point of most discussions of the law is compliance, since the purpose of creating laws and empowering legal authorities is to establish and maintain social order by regulating public behavior. (…)*
>
> *At the time that* Why People Obey the Law *was written, the conception of the relationship between community residents and legal authorities was a reactive one, with obedience to legal rules viewed as the key behavior that legal authorites wanted from those in the community. Since that time it has been recognized that authorities need the more active cooperation of those in the community.*
>
> *Tom R. TYLER*– Afterword *to* Why People Obey the Law *(2006 edition)*

> *Good policy analysis is not about choosing between the free market and government regulation. Nor is it simply deciding what the law should proscribe. (…)*
>
> *Participants on both sides frame the deregulation debate as a kind of "Live Free or Die" policy choice. Even lovers of liberty might reasonably ask whether third alternatives do not exist.*
>
> *Ian AYRES and John BRAITHWAITE* – Responsive Regulation (*1992)*

After having sketched out the historical emergence and evolution of the regulatory inspection function (or at least of *some* regulatory inspection functions), and before we consider current examples of purported "risk-based inspections" (and compare them to other practices), it is necessary to summarize and discuss the theoretical underpinnings and research findings that can shed light on both "regulatory inspections" and "risk".

We will consider prior research mostly from three perspectives. First, an introductory section where we will summarize perspectives on the uses and appropriateness of regulation, and on the question of its costs and effectiveness[271], since risk-based inspections are touted as a way to improve both. Before we look at what data can indicate of practical results, it is thus needed to look at the context against which risk-based inspection reforms are implemented – how relevant regulation is both to economic issues, and to its purported social welfare goals. Second, we will look at theories seeking to account for regulatory compliance, and how

[271] Considering here not only "regulation" as a whole, but also to some extent specific *regulatory instruments* – the distinction being here that "regulation" is a set of rules, to which economic operators are subject, and "regulatory instruments" are specific procedures and processes through which these rules are administered, implemented, enforced etc.

well they seem to fare in experimental research. Indeed, given that the primary justification given for the existence of inspections is generally the aim to increase compliance, understanding better what drives compliance is vital to attempting to better assess inspections' effectiveness, and the ways in which it might be improved. Discussing compliance visions will also enable us to briefly touch on the question of regulatory discretion, which is a fundamental element of risk-based approaches (and one that is, at times, hotly debated). Third, and finally, we will attempt to summarize at least some of the considerable amount of research that has developed on the interaction of risk and regulation. While we may not purport to be exhaustive on this count, these insights will be crucial to put risk-based inspections in perspective, and help clarify the meaning of risk and challenges associated with risk-based approaches.

3.1. Regulation: uses, costs and effects – a brief overview

a. The uses and abuses of regulation – introduction

The very word "regulation" has a wealth of meanings, and is far from uncontroversial. It can be (both in English and its different translations) understood to mean in a legal sense any sort of secondary legislation (decrees or other norms issued by the executive), or at the other extreme a complex system ensuring a cybernetic equilibrium, be it is in economics or social science. Over the past couple of decades, the word has acquired also a specific use in relation with economic activities, but even this field sees several competing meanings – with "regulation" either used for the oversight and control of prices and services imposed on monopoly or quasi-monopoly privatized (or quasi-privatized) utilities (and providers of fundamental consumer services), or for the entire set of rules (technical, fiscal, related to starting or closing an activity, etc.) applicable to economic operators.

It is this latter sense, which is sometimes called "non-economic regulation" (to distinguish it from regulation of utilities etc.) that is relevant to our research. Within this field, we in fact focus mostly on a specific subset of regulations that relate to safety and health in the broadest sense, including environmental protection, and the protection of other public interests – including product market regulations, as well as regulations relating to the construction and operation of business premises[272]. While the use of "regulation" in a specific sense (or rather, at least two specific senses) in the economic sphere has gained international acceptance[273], it is not necessarily uncontroversial. Why, some ask, should laws, decrees and other norms that apply to businesses (or to private citizens acting in an economic capacity, e.g. as "sole traders") be treated differently from other laws, suggesting in some ways that they are "less legitimate" or "less mandatory" than other rules[274]?

---

[272] Unfortunately, no satisfactory single term currently exists to cover this sub-set of business regulations – "technical regulations" has a specific WTO TBT meaning, "health and safety" is often understood to mean only/mostly "occupational safety and health", etc. Moreover, some of these regulations do not directly relate to *safety*, but to other public interests, e.g. consumer information etc.

[273] See e.g. different OECD publications, where "regulation" and "regulators" are understood in subtly different ways: in the *OECD Best Practice Principles for the Governance of Regulators* (2014, available at: http://www.oecd.org/gov/regulatory-policy/governance-regulators.htm) and the *OECD Best Practice Principles for Regulatory Enforcement and Inspections* (2014, available at: http://www.oecd.org/gov/regulatory-policy/enforcement-inspections.htm). Interestingly, the OECD treats the use of "regulation" in the specific economic sense as fully obvious, but very rarely attempts to define it. One such attempt is in an early document in the OECD's "Regulatory Policy" workstream, the 1995 *Recommendation of the Council on Improving the Quality of Government Regulation*, which refers to the "framework of responsibilities and constraints established by government regulation" – in this (very broad) meaning, "regulation" covers all the rules (creating obligations or prohibitions) that apply to economic operators.

[274] See e.g. the views of Carson as summarized by Hawkins (2002): "prosecution as a last resort in Victorian times can be seen as evidence of a process of 'convetionalization' of occupational safety and health offences. Hist contention is that such offences were suffused with a sense of ambiguity which led to matters formally enacted as criminal becoming regarded as merely quasi-criminal and

Engaging in depth with this debate would take us far beyond the scope of this research, but some level of clarification is nonetheless needed to give our work a sound basis. First, we will try and articulate very briefly why it can be held as legitimate to handle "regulation" distinctly from other legislation. Second, we will summarize some of the prevailing views on why and how regulation should be used. Finally, we will attempt to sketch out why, in our view, the attempt to oppose "smarter regulation" because it would show undue leniency towards businesses is misguided, but rather the principles and tools of "smarter regulation" should also be used in matters that do not pertain to businesses but to citizens' private lives, as they are sound ways to make public policy in general more proportionate and effective.

i. Questions around the legitimacy of treating "regulation" specifically

Historical introduction – economic freedom and regulation

When considering the legitimacy and appropriateness of treating "regulation" as a distinct field, the historical perspective cannot be avoided. The significant restrictive rules that affected economic activity in pre-modern times, including various duties, tolls and levies, monopolies, restriction on entry, product-related rules etc., all corresponded to a situation where economic activity[275] was regarded as part of a broader social order, a collective undertaking where each member of society had to carry tasks according to its assigned place. In the medieval tri-partite vision, alongside those in charge of prayer and of fighting, were the many assigned to labour – and, among them, each had his or her role. Movement was very much discouraged, as a form of challenge against the God-assigned order, and the established powers, both spiritual and secular. In such a world, rules restricting certain trades to guild members, setting down exactly how products should be manufactured, limiting trade etc. were but manifestations of the social order, as necessary and as little disputed as the rules of monastic orders. They cemented the cohesion of the community, and the various duties and levies both ensured the funding of the praying and fighting orders, and protected local producers against competition, again fully in line with the broader social vision[276]. The gradual changes in world view, social order and economic structures that took place over the 15th to 18th centuries brought about a complete reversal[277], with the notion of *freedom*, specifically of *individual freedom* – and, alongside the political one, of *economic freedom*. In the new social order, such as it emerged in France and Britain after the French Revolution and the more than two decades of wars that ensued, political freedom was far from always ensured, but economic one was secured to a large extent. "*Laisser faire, laisser passer*" became, if not always the norm in practice, at least the position that best reflected dominant ideology.

In such a context, regulation of economic activities can of course exist as a "left-over" of the previous social order (e.g. the persistence of regulations on certain professions such as notaries in France, even after the Revolution), or can arise as the result of conflicting values (e.g. the demand for more social justice, or concerns about keeping "order"), but it can also be developed in a way that is internally coherent with the primacy of individual (economic) freedom. Indeed, as the 1789 *Déclaration des Droits de l'Homme et du Citoyen* puts it, "freedom consists in being able to do everything that does not harm others: thus, the exercise of each man's natural rights has no other bounds than those that ensure that other Members of Society can enjoy these same rights" (article 4). This means, in an economic perspective, that regulation that limits economic freedom

---

not as 'real crime' at all" – a situation "revealed, for example, in the fact that offenders were not dealt with as part of the usual criminal justice system, but by regulatory bodies" (p. 19).

[275] Which, of course, was not considered under this name at all. The words "economy" and "economics" in their modern meaning only started being used in the late 18th century.

[276] See Duby (1978), which remains the fundamental work on this topic. Many other works have covered this topic since then, e.g. Arnoux (2012) – but Duby's work remains valid.

[277] See in particular Gauchet (1985), but also Mercier (1960), Muchembled (1988) *et al.*

can be legitimate when the effects of economic activity would harm the freedom of other people, including by endangering their health, or affecting their property (as, in the view of the *Déclaration'*s authors, there can be no freedom without safety of body and property). Within this framework, considering "regulations" as *distinct* from other fields of legislation is logical – because it limits specific (economic) freedoms. Such specificity, however, is not different in nature from that which should apply to other areas of legislation that limit other fundamental freedoms (e.g. press and media law). Thus, the internal coherence of an individualistic and liberal world view (and legal order) makes it legitimate to consider economic operations' regulation specifically, but not more so than a number of other legal domains, which also impinge on freedoms.

Clearly, that "regulation" has come to be named and handled in a specific way, different from other "freedom infringing" areas of legislation, is due to the centrality of the issues it impacts on for modern societies: wealth, distribution of income, labour relations, economic and political power, growth and employment etc. This manifests itself both in terms of regulatory capture due to the power of influence wielded by economic operators, but also in terms of regulations developed with a specific "anti-business" intent, supported by political and social forces critical of the existing economic order. Regulation is, thus, a particularly *contentious* area of legislation.

In line with its contentious nature, regulation has been (and still is) criticized from a variety of corners, with the different perspectives reflecting to some extent ideological preconceptions, but also to a large extent the diversity of regulatory questions, and the complexity of regulatory interactions. Hawkins (2002) provides a very condensed summary of what he calls "the debate about command and control regulation" (pp. 13-15). Because this summary is both clear and comprehensive, we will just refer readers back to it for details, and present only the key elements here. First, while "command and control regulation is generally justified in instrumental terms" (Baldwin 1995 *et al.*), its effectiveness is often far from optimal, and many authors have linked this to the "capture" of the regulators (Bernstein 1955 *et al.*), to a regulatory life-cycle where the "energy of the regulatory body is sapped" (*ibid.*), or to the interplay of "interest groups" (Posner 1974). Others have suggested that the problem may be in the nature of command and control itself, that tend to lead to costly, inefficient, short-term solutions (Sinclair 1998), to complexity, rigidity, costs and delays (Bardach and Kagan 1982). Designing "perfect" or "optimal" rules seems impossible, and rules tend to fail on both sides, creating high costs with limited effectiveness (Baldwin 1995). As for the practice, negotiation is often the rule in enforcement (Hawkins 2002, Hutter 1997), with some authors lamenting the lack of more vigorous enforcement (Tombs and Whyte 2008, Pearce and Tombs 2009). Others suggest that more "responsive" or "smarter" enforcement can lead to regulation that is more efficient and more effective (Ayres and Braithwaite 1992, Gunningham and Grabosky 1998).

What is of great interest here, in our view, is that in fact these challenges (and potential solutions) are in no way exclusive to regulation of economic operators (or of "businesses"). They are, in fact, to varying extents, applicable to any set of rules that have an *instrumental* purpose[278] - i.e., they tend to be far less effective than their proponents envisioned, create important costs and side-effects, be difficult to enforce – and it may be that "smarter" enforcement methods allow to improve their effectiveness. The specificity of regulation may well reside primarily in the fact that, because of the centrality of economic issues in our societies and of the strength of the different interest groups involved, a real discussion has arisen around them, including on the question of their implementation and enforcement, which may well hold lessons for other areas of legislation.

Distinguishing between different categories of norms, and different uses of legislation

[278] See Hawkins (2002) pp. 3-13 for a discussion of "instrumental" vs. other (in particular "symbolic") uses of the law.

The question of whether it may be acceptable to differentiate enforcement approaches (inspections frequency, enforcement decisions etc.) based on the level of risk (and other factors) is tied to the nature and status of the legal norms being enforced. This, in turn, relates to the possible distinction between different types of laws and norms – and between different uses of legislation.

Types of norms and levels of obligation

One way to attempt and make sense of the distinction between regulation of economic activities and other parts of legislation is to consider the difference between several types of laws or norms – in the perspective of the old question of a "natural law", and to the possibility (or lack thereof) to distinguish between norms that would correspond to an "overlap" between natural and positive law, and other norms that would only belong to positive law, but not carry greater weight, i.e. obligate but not "in the fullest sense" (Finnis 1980 quoted in Hemma 2015).

The idea of a "natural law" is as problematic as it is old, and long-debated. It can be understood to have a huge variety of meanings, and is tightly linked to a series of religious, philosophical or ideological perspectives (Goyard-Fabre 2002, pp. 7-8). Recent controversies and judicial decisions in the United States around homosexual marriage, which featured references to "natural law" among opponents, and reference to "fundamental human rights" among supporters, show the difficulties and ambiguities that abound in this notion. Nonetheless, just as the philosophical discussion around the idea of a natural law should not be avoided and can yield real fruits (*ibid.,* pp. 14-15), the distinctions it enables to introduce can shed some light to our topic.

Rather than going back all the way to Aquinas and different interpretations of classical natural law theory, we will draw on a few modern authors, whose ideas bear clear relevance to this research. First, as indicated above, Finnis distinguishes between "obligation" and "full obligation": the "essential function of law is to provide a justification for state coercion (…). Accordingly, an unjust law can be legally valid, but it cannot provide an adequate justification for use of the state coercive power and is hence not obligatory in the fullest sense" (Hemma 2015). This view does not really *conflict* with legal positivism (Finnis does not challenge the validity of positive laws), but introduces a nuance into the strength of the obligation they impose. Laws that correspond to an overlap between fundamental moral norms and positive law have, in this view, a power of obligation "in the fullest sense".

Second, Dworkin considers "that there are some legal standards the authority of which cannot be explained in terms of social facts. In deciding hard cases, for example, judges often invoke moral principles that (…) do not derive their legal authority from the social criteria of legality contained in a rule of recognition" (Dworkin 1977, p. 40, quoted in Hemma 2015). Dworkin uses as an example the famous *Riggs v. Palmer* 1889 decision by the Court of Appeals of New York, wherein the Court decided that a murderer could not benefit from his victim's will, even though there was no positive law to back their decision – drawing on "a requirement of fundamental fairness that figures into the best moral justification for a society's legal practices considered as a whole" (*ibid.*). Further to this, and in the same perspective, Dworkin introduces a fundamental distinction between "two kinds of legal argument. Arguments of policy "justify a political decision by showing that the decision advances or protects some collective goal of the community as a whole" (Dworkin 1977, 82). In contrast, arguments of principle "justify a political decision by showing that the decision respects or secures some individual or group right" (Dworkin 1977, 82). On Dworkin's view, while the legislature may legitimately enact laws that are justified by arguments of policy, courts may not pursue such arguments in deciding cases. For a consequentialist argument of policy can never provide an adequate justification for deciding in favor of one party's claim of right and against another party's claim of right. An appeal to a pre-existing right, according to Dworkin, can ultimately be justified only by an argument of principle" (Hemma 2015). This distinction is of great importance for us, in that a large part of the norms subsumed under the "regulation" moniker are clearly expressions of *policy choices*, but not of fundamental *rights and principles*.

Third, Fuller's vision of "procedural morality" in law posits that "law's essential function is to "achiev[e] [social] order through subjecting people's conduct to the guidance of general rules by which they may themselves orient their behavior" (Fuller 1965, 657)" and this "implies that nothing can count as law unless it is capable of performing law's essential function of guiding behavior" (Hemma 2015). In order to perform this function, "a system of rules must satisfy the following principles: (P1) the rules must be expressed in general terms; (P2) the rules must be publicly promulgated; (P3) the rules must be prospective in effect; (P4) the rules must be expressed in understandable terms; (P5) the rules must be consistent with one another; (P6) the rules must not require conduct beyond the powers of the affected parties; (P7) the rules must not be changed so frequently that the subject cannot rely on them; and (P8) the rules must be administered in a manner consistent with their wording. On Fuller's view, no system of rules that fails minimally to satisfy these principles of legality can achieve law's essential purpose of achieving social order through the use of rules that guide behavior." (*ibid.*).

Connecting these views to our field of research is easy, and enlightening. First, most matters covered by regulation simply do not relate to fundamental issues of morality (whichever way, and on whichever basis one construes them), and thus fail in Finnis's perspective to "fully obligate" (they do obligate, but in a "lesser" way). Second, most norms pertaining to regulation are adopted in order to advance policy choices, and do not relate to fundamental rights and principles – and thus fail to carry the same weight, even though they are legally binding. Third, the principles identified by Fuller as necessary for the law to achieve its purpose form the foundation of many "better regulation" or "smart regulation" principles, showing the link from these newer approaches to longer-standing visions of good legal practice. These (and in particular the first two points) form important theoretical justifications for practices that we will consider further in this research, and which involve a level of discretion in the enforcement of regulation[279].

Nor are these purely theoretical, but rather jurisprudential practice shows the relevance of these distinctions. In international law, for instance, the notion of *Jus Cogens* (peremptory norm) refers to norms for which no exception or variation is admitted, for their moral strength (viewed as applying to all humanity, throughout moral systems) gives them particular weight[280]. By contrast, other norms arise through convention, and do not carry the same peremptory strength. The whole tradition and practice of Common Law is likewise built on the idea that some fundamental practices can be identified, and built upon, even in the absence of a positive norm. Such idea is not absent from Civil Law countries either: in France, the "*principes généraux du droit*" (general principles of law), which apply primarily (but not only) to administrative law, can lead to courts ruling against administrative norms and decisions based on principles rather than positive law[281].

From this perspective, as a result, it appears legitimate to challenge the view, which we have seen held in many countries, that risk-focus and risk-proportionality would be somehow illegitimate because they would conflict with the absolute obligation created by law, and the absolute duty for the executive to enforce it.

---

[279] The inspiration for this section was provided by a presentation by Donald Macrae at the International Seminar on Regulatory Discretion held in The Hague in December 2013 – in which he presented a vision of a "hierarchy of norms" – the most fundamental ones expressing *values* (and carrying the most weight, being the most "peremptory"), a second category being the foundation of *order* (e.g. driving rules), and thus having to be strictly complied with in spite of them being purely conventional – and a third category corresponding to the bulk of regulation, and expressing *policy.* We have tried here to provide a theoretical underpinning for this distinction which, in our experience, is extremely valid in practice. The presentation can be accessed at: http://www.ial-online.org/uploads/2014/01/The-Hague-131205-session-2-presentation-1-Macrae1.pdf.

[280] For illustrations, and discussions of the effects of what the author sees as "excessive" application of the criminal law to regulatory issues, cf. Malcolm (2014 a) – "Unlike *malum in se* offenses, most criminal regulations do not prohibit morally indefensible conduct. Regulations allow conduct, but they circumscribe—often in ways that are very hard for the non-expert to understand—when, where, how, how often, and by whom certain conduct can be done" (p. 1).

[281] These principles are "identified" drawing on "ideological conceptions of the national consciousness" and a "mass" of national, international and other texts (Frier and Petit, quoted in Tifine 2010, 2nd part, chapter 1, section IV).

Rather, as these authors suggest, there are meaningful distinctions to be made between different types of norms, which carry different levels of obligation.

Legislation and regulation – the different uses of law

Another important distinction is between the different purposes of *legislation* and *regulation*, which are essentially distinct in spite of their important overlaps – and between the different uses of law. As Voermans (forthcoming) puts it, "with 'legislation' we mean the authoritative, and constitutionally controlled form in which law is cast and the procedure leading up to the enactment of it (the decision). With regulation we mean a public intervention in a market or in society. (…) Legislation and regulation coincide in a lot of instances. A lot of regulation is cast in the form of legislation. (…) But not all regulation needs to be cast in the form of legislation (…), and not all legislation is regulation". While the distinction is essential, and shows that there is no full overlap between the two notions, what matters even more to us here is that the two have a different *focus* and *purpose*. To quote Voermans again, "the focus of and underlying notion of regulation (…) is on government intervention in markets, i.e. on acts private actors cannot perform with private capital, on interventions beyond regular market mechanisms (…) Legislation on the other hand focuses not primarily on markets but – to use a big word – on the human psyche, especially morale and social relations: the oughts of our existence."

The scope of legislation is thus much broader than that of regulation – which, to the extent that it is cast in the form of legislation, can be seen as a specific subset of the broader field of all legislation. Beyond their different focus, the two also have different goals. Regulation "predominantly functions as a market intervention aiming for a correction" (Voermans, forthcoming), and this holds true regardless of whether one considers *normative* or *positive* theories of regulation. Normative ones will consider the instances in which regulation could be seen as appropriate from an economic perspective, in particular to address market failures and inefficiencies, but also issues of distribution, fairness etc. in some instances (Veljanovski 2010, pp. 22-24). Positive ones will look at how regulation is produced in practice – interest groups at play, effects on wealth transfers between different groups, etc. (*ibid*., pp. 23-26). In all cases, the focus is market relationships and economic issues. By contrast, legislation "serves other and broader functions": it "provides both the basis and the framework for government action", "works as a safeguard against government action by enshrining rights and obligations" and provides "legal certainty". It also can "serve as an instrument to further government policies (instrumental function)", "offers the basic framework for the operation of a bureaucracy" and "communicates and reaffirms public morals, values and public goods (symbolic function)" (Voermans, forthcoming).

In spite of the apparently clear differences, there exists a tension because legislation increasingly has been used over the past century and a half to make "continual improvements in the life of the community by means of explicit legal innovations" rather than (as was hitherto its most common function) being predominantly a "benign instrument of codification through which hitherto scattered and inaccessible common law could be systematized and made accessible for everyone" (*ibid.*). To the extent that legislation is an instrument for polciy objectives, and that some of these policy objectives affect economic issues, there is a significant overlap with regulation. While regulation primarily focuses on affecting economic activities to achieve specific goals, legislation more broadly seen has a number of other fundamental roles – expressing moral values, and ensuring the functioning of the constitutional order. As Voermans (*ibid*.) puts it, "from a constitutional point of view (and the symbolic function which is closely related to it) the only right measure for the quality of legislation is its ability to express law" and "the extent to which the criteria, emanating from constitutional principles, are met" – whereas regulation treats legislation as a means to other ends, and assessing its quality thus requires to take an *instrumental* perspective.

To the extent that a significant part of regulation is enacted by legislation, and that the purposes and criteria are fundamentally different, there can thus be cases where different perspectives will result in conflicting views on enforcement. Reducing legislation to an instrumental perspective is inherently problematic – Voermans (*ibid*.), referring to Tamanaha (2006), reminds that "instrumentalism may in the end undermine important social and symbolic functions pertaining to legislation". In an interesting twist, there are strong reasons to believe that the more "social and symbolic functions" of legislation themselves have important economic value – for instance, "European legislation also creates trust, security, legal protection and all kinds of other, more or less imponderable, benefits for the internal market" meaning that the "pricing" of costs and benefits may be "extremely difficult" (Voermans, forthcoming).

There is thus no easy solution to our issue – from an instrumental perspective (befitting "regulation"), it may make sense for enforcement to be responsive and risk-proportionate, but this may conflict with other values expressed by legislation. Voermans (*ibid.*) reminds rightly about the relevance of a political perspective: because "Better regulation" and "Better lawmaking" policies are "essentially political programmes resting on political perceptions as to the overriding values of legislation and regulation", their effect and success has to be "weighed politically". In other words, there can be no politically neutral consensus on the right approach, but rather one can look at the adequacy of a given approach (or programme) to a clearly stated political objective.

A last point of note is that importance of *trust* – Voermans (*ibid.*) repeatedly emphasizes how essential the function of *building trust* is for legislation. Not only was this the key role of legislation in enabling markets before regulation with specific "transformative" goals came about – but it is a role that has remained crucial. In fact, enabling trust is one of the fundamental functions of modern regulation, as we have seen above e.g. with respect to food safety legislation. A criterion that may therefore be common to both a "regulatory" and a "legislative" perspective is whether enforcement practices actually are effective at reinforcing trust between market actors, or not.

ii. Justifications of regulation – why, when and how to regulate

If regulation is indeed primarily the expression of *policy preferences*, and not of fundamental principles or rights (though it is the latter in *some* cases), some guiding principles are needed in order to define why, when and how to regulate. In such a perspective, regulation is a *policy tool* – as any tool, it is not all-purpose or "one size fits all", and can produce damage as well as positive results. Thus, such principles are essential. It is fair to say that, at least for authors who place themselves within the framework of a broadly "liberal" market economy, the most broadly accepted foundation for regulation is what is called *market failure*. Even considering authors that advance a different view of society and the economy, the following principles may remain applicable insofar as they also relate to the best choice of instruments, and not only to the goals being pursued (even in a radically redistributive perspective, for instance, regulation may not be the best option, as compared to taxation and spending, for instance).

Anthony Ogus, in *Regulation. Legal Form and Economic Theory* (1994), did far more than give a specific account of how regulation had evolved and acquired more prominence as a public policy instrument in the 1980s – he attempted to give a comprehensive account of regulation's foundations, purposes and forms from a *normative* perspective. This perspective can be complemented, in particular for a concise summary of the *positive* perspective on regulation, with Veljanosvki (2010).

Understanding the role and limitations of regulation in a market economy is crucial in order to the consideration of inspections and enforcement – because they cannot be considered fully separately from the

rules they are meant to enforce. Inspectorates define priorities, give guidance and instructions to their officials, and in many cases adopt and publish guidance documents for the public, or even secondary legislation. We have observed in many cases how these were often based on a vision of regulation as an all-purpose instrument, that could be used in any circumstance, for any type of problem, and was expected to be effective in all of them (and, consequently, inspections and enforcement would likewise be appropriate to solve this problem). By contrast, a more precise and limited vision of what regulation can really achieve, and of when it is appropriate, is fundamental to define priorities and methods in a more targeted, focused and differentiated way, which is what risk-based and compliance-focused approaches generally seek to achieve. In this perspective, we will thus briefly summarize some of the key normative perspectives on the proper role and instruments of regulation.

Why and when to regulate

In a market-based context, regulation comes as an *exception* to free economic activity. In an "ideal" market, parties should be left fully free to contract – but the need for regulation arises primarily from "imperfections" in the market, what is broadly termed "market failures", which mostly arise when "negative externalities" (negative effects of economic activities, beside and beyond their main purpose, affecting third-parties) are significant (and not addressed), or when "transaction costs" (costs needed for information gathering, negotiation, transaction) are too high (Ogus 1994 pp. 17-19[282]). In theory, and again in an ideal (and clearly unreal) market setting, negative externalities could be dealt with through private contracting[283] - in practice, however, this is often impossible, either because transaction costs are too high, because some externalities are not "priced" (some goods are entirely free and there are major problems involved in "privatizing" them to allow for contracting to resolve externalities – e.g. this is the case of air), because the benefits are highly concentrated and the harm diffuse (making collective action unlikely and costly), to name just the main problems (Ogus 1994 pp. 19-22).

Situations where negative externalities of economic activities are significant, and where private contracting cannot provide an adequate response, are at the root of most of the regulations for which we consider inspections and enforcement in this work. In some cases, it is possible to avoid using "command-and-control" regulations by relying on tort law (Ogus pp. 20-21) and private law more generally, but "the courts have jurisdiction to enforce rights only *ex post*", meaning "after the damage has been inflicted" – and in some cases the infringer may "avoid the sanction by insolvency" (*ibid.,* p. 28). In any case, in many cases relying on private law will be inadequate because potential plaintiffs "will only seek to enforce rights where the expected benefits exceed the expected costs" and "thus externalities which affect large numbers but which impose only a small loss on each (…) will not be 'internalized' by private law instruments and serious misallocations will remain uncorrected" (*ibid.,* p. 27). This is not mentioning the serious problems arising when the right-holders are in a situation where they are ill-equipped to avail themselves of the judiciary (poverty, lack of legal literacy, etc.) – and assuming an unbiased judiciary, of course. In short, there are situations where "market failure" is

[282] See also *ibid.* pp. 41-42 on "coordination problems" i.e. issues where in principle negotiated agreement would be possible, but the number of actors and interactions makes it absurd, e.g. the driving code. While the driving code rules to a large extent are purely conventional (e.g. whether to drive on the left or right side of the road), having each pair of drivers negotiate them is simply impossible (and even absurd). A regulatory intervention is far more "optimal", and in fact clearly necessary (and this corresponds to what we described above as essential conventions allowing the proper functioning of society).

[283] And this is what radical (right-wing) libertarians like Nozick (or, earlier, Hayek) would advocate: no regulation, only private contracts. We will not discuss here the many problems that plague such views, but essentially the problem of transaction costs is the first that, even in a purely market-oriented worldview, makes the full reliance on private law inadequate. Akerlof (1970) provides a perfect example of why information asymmetries make it in practice impossible for many markets to function properly absent any regulatory intervention.

accompanied by "private law failure", which builds "on public interest grounds a prima facie case for regulatory intervention" (*ibid.*, p. 28).

In practice, there frequently are problems caused by unaddressed negative externalities and high transaction costs – because the assumptions for perfect functioning of the market are rarely met – these include fully rational, "utility maximizing behaviour" by all market actors, sufficient information for all actors "to make utility-maximizing choices"[284], absence (or full correction by private law mechanisms) of negative externalities, and fully "competitive markets" (*ibid.,* p. 24). In practice, these conditions are generally only partially and imperfectly met, at best. Thus, regulation can be needed, and can aim at addressing any or all of these problems and imperfections – prohibiting or constraining operations that create significant negative externalities, reducing transaction costs by establishing uniform requirements for products[285]. Regulation can also focus on specific market imperfections, e.g. mandate the disclosure of specific information in a standardized way, "nudge" people towards more utility-maximizing behaviour[286], or intervene to limit the power of dominant market actors[287].

Thus, overall, regulation can be justified in a such a system when, absent regulatory intervention, there would be a problem of inefficient allocation of resources – for which, following Ogus, we would adopt the Kaldor-Hicks criterion rather than the Pareto one[288]. Addressing market failures and private law failures, "inefficiencies" in economic terms, "infringements of rights" (including the right to life, in some cases), all require regulation, at least in some cases. But it does not follow that regulation always works as intended, or that the form and tools of regulation are indifferent. Ogus of course discusses the different ways in which regulation can fail its objectives, be driven by private interests from the onset, or "captured" during implementation (see in particular *ibid.* pp. 55-58 on "regulatory failure"). We have already discussed this question above, and will now focus on how to select appropriate instruments for regulation.

How to regulate

The types of regulation that are most commonly controlled through inspections are mandatory technical norms (that Ogus calls "standards", a name which we avoid here because of its polysemy[289]), as well as

[284] See Ogus (1994) pp. 38-41 on the problems of information often being limited, imperfect, costly or impossible to process.

[285] Of any kind: these can be uniform requirements for physical products (e.g. food),but also for financial ones (loans or insurance contracts), rental agreements etc.

[286] See Sunstein and Thaler 2008. While it is frequently understood that the "nudge" approach is an *alternative to regulation,* it is in fact often better understood as an alternative to "command and control" regulation. Behavioural economics insights are used to design regulatory requirements that require specific ways to disclose or present information, mandate some default options, etc. For more on the use of behavioural economics in regulatory policy see Lunn (2014) report to the OECD, and Alemanno and Sibony (2015). In particular, see Lunn pp. 39-41 on the application of behavioural economics insights to "regulatory delivery" (including inspections and enforcement).

[287] Be it in positions of monopoly/oligopoly, or monopsony/oligopsony – in practice, most regulation focuses on (quasi-)monopolies linked to natural resources, utilities etc.

[288] See Ogus 1994 pp. 24-25 – whereas a Pareto distribution is efficient if it is impossible to make any change without making at least one person worse off, and thus prohibits improvements that would benefit the vast majority if even the smallest minority stands to lose from them, the Kaldor-Hicks test allows for *compensation.* In this meaning, a policy is efficient if the overall gains it produces are sufficient to *potentially* allow to fully compensate all the losers and *still* produce an aggregate benefit. It is easy to understand that these two definitions of "efficiency" lead to radically different policy perspectives (and tend to correspond to radically different political sides, as well).

[289] The word *standards* can have at least three major meanings. In its *technical sense* (used in the WTO TBT agreement for instance), it is a document that provides requirements, specifications, guidelines or characteristics that can be used consistently to ensure that materials, products, processes and services are fit for their purpose . Standards are *voluntary* in nature, and developed by institutions (national, regional – CEN, CENELEC etc. – or international – ISO) that are normally acting on behalf of stakeholders (particularly businesses) and not of public authorities (which, however, often provide some funding to standardization bodies). On the other hand, in countries where many standards are mandatory, the standardization body is often a state agency. In its *vernacular meaning,* a "standard" is a norm, convention or requirement of any kind – but also can mean (as when one writes "the highest standards") the

information obligations[290]. Information obligations can cover a variety of issues and fields, and include e.g. mandatory price disclosure (Ogus 1994 pp. 126-128), weights and measures[291] and requirements on display of quantity (*ibid.,* pp. 130-132), and rules on "identity and quality disclosure", i.e name, description and composition of products[292] (*ibid.,* pp. 132-138) and warnings and instructions about use of the product being sold (*ibid.,* pp. 141-144).

Information requirements are not cost-free, and there can be a tendency on the side of regulators to impose too many of them, because they are *mutatis mutandis* significantly less restrictive than mandatory technical norms. Information requirements still allow economic operators to produce and market goods pretty much as they decide to, provided that they comply with rules in terms of labelling and other information. Nonetheless, the way they are worded and controlled can result in higher or lower constraints and costs for economic operators (and, in turn, in different economic effects).

In spite of the importance of information obligations, and of their pervasiveness, they have been the subject of relatively less study (and debates) than mandatory technical norms, possibly because of the latter's more "reassuring" character (what may be hazardous is forbidden, rather than just carrying warnings) and of the greater economic distortion they can impose (direct restriction on the possibility to bring products to market).

Mandatory technical norms are one of a gradient of regulatory interventions, ranging from the least to the most restrictive. Ogus ranks such intervention types (*ibid.*, p. 151) with information requirements as the least restrictive, prior approval as the most restrictive[293], and "standards" (mandatory technical norms) in between. Ogus further differentiates between "target", "performance" and "specification" standards (mandatory technical norms), and this is a distinction that is very important for inspection practices. We will not summarize here the detailed discussion of cost-benefit aspects of different types of interventions, and of cost-benefit analysis models (*ibid.*, pp. 155-165), on which there is considerable literature[294]. The distinction between "target", "performance" and "specification" norms is, however, central for inspection work (*ibid.,* pp. 166-171).

While "target" norms "render unlawful the causing of certain harms" (p. 166), they do not specify how an economic operator should conduct its activities, nor do they deal with "intermediary outcomes", which may arise between the activities and the harms that the norms aim at preventing. Thus, while they allow the greatest flexibility in economic operations, and thus could theoretically be the ones that impose the least

---

way in which something is done or executed, regardless of whether this is codified or not. There is often the assumption that when "standards" are spoken of, then "high standards" are expected, and that more or less automtically "standards" are "a good things" (hence: more standards are better). This has implications in policy discussions, where there is often an automatic bias for "more standards". Finally, *regulatory standards* (the meaning Ogus uses) refer to mandatory technical norms (applying to anything from hygiene to occupational health, fire safety to environment). Given the potential for confusion, we prefer not to use the word at all here, and refer to mandatory technical norms.

[290] Other forms of regulation covered by Ogus, e.g. prior approval (licensing, permits etc.) and economic instruments (incentives etc.) also frequently involve different forms of inspections – they are nonetheless less "central" in the work of inspecting institutions.

[291] Which, as we have seen above, belong to the oldest areas of government regulation, as well as inspections.

[292] Such requirements can *mandate* that specific information be given, and/or regulate *when and how* the use of certain names, descriptions or claims can be allowed.

[293] In fact, information requirements could be further disaggregated between different types: disclosure rules vs. restrictions on the use of certain names and descriptions, for instance. Likewise, "prior approval" can cover a number of situations with varying degrees of restrictions, requirements, short or long procedures, need to obtain other "prior prior approvals" etc. And, of course, these requirements can be *combined*: the same business operator and product can be subject to prior approval, and then mandatory technical norms, and information requirements in addition (in fact, this is generally the case that "stronger" requirements come *on top* of "weaker" ones).

[294] See e.g. Radaelli and Dunlop (in press), and of course a number of previous publications e.g. by OECD. The question of cost and benefits is of course connected to the question of inspections and enforcement, in that the potential costs and effectiveness of inspections and enforcement should be considered when conducting RIA (or any other form of impact assessment). In practice, they often are not, or insufficiently, taken into account.

burden on the economy (and least distort allocation of resources), they often create difficulties. Depending on the level of discretion and authority granted to the regulatory agency in charge of enforcing these norms, the difficulties may be more on the side of economic operators, or of regulators. Where regulators have limited discretion and their authority is subject to strict judicial review, it may be difficult for them to enforce such norms, because of the difficulty to prove a causal relationship between specific economic activities and the harms covered by the norms, and/or the time lag may be too long for effective prevention. At the same time, such norms can also be very problematic for economic operators, because of high uncertainty ("the information costs to the firm on determining what quality of performance will ensure compliance" may be high because of uncertain causalities, third party activities also having an effect, etc.). Thus, while such standards are attractive at first glance from an economic efficiency perspective, and can be particularly conducive to innovation and technological flexibility, they carry potentially important costs and operational difficulties.

"Performance" norms are somewhat less "uncertain", while leaving a significant room for flexibility. They impose prescriptions (e.g. maximum level of certain emissions) on the direct outcomes of economic activities, while leaving the specifics of the operations open. Such norms have the advantage of providing regulators with more directly verifiable (and enforceable) indicators, and of burdening economic operators with less uncertainty. They also leave a fair amount of room for technological innovation (though less than "target" norms). However, because they focus on intermediate outcomes and not the final harms that the regulation aims at reducing, they can fail (partly or fully) in preventing or reducing such harms, if the causality between regulated outcomes and harms is less strong than anticipated, and/or any unexpected effects take place (involving third parties, side effects etc.).

Finally, "specification" norms directly impose how certain economic activities should take place, which materials, products, processes, methods are allowed, which ones forbidden, etc. The relationship between "specification" norms and the harms they are supposed to prevent is indirect, and it can often happen that the norms cover a number of issues but fail to address the harm because some critical issues were left out (because of limited knowledge, or poor design, etc.). To address one given issue, "it is often necessary to lay down a series of specification standards" (*ibid.*, p. 167), resulting in a large number of norms, but with compliance requirements being clearer (and enforcement simpler) than with "target" or "performance" norms. While "specification" norms thus bring greater certainty and predictability (and may make deterrence easier and stronger), they have "significant disadvantages", in particular inducing high "technological rigidity" and making it more difficult to introduce new techniques, methods, processes (even when these would actually improve performance in terms of harm reduction). They also often, as indicated, fail at preventing harm because they do not address it comprehensively, but rather target only some of the precursors of harm, and may miss some critical ones. Such highly detailed and prescriptive norms are also, in most cases, the oldest type of norms[295], and the most widespread[296].

From an inspector's perspective, "specification" norms hold much appeal: they are clear and unequivocal, lend themselves to relatively easy enforcement decisions, make control work easier and faster, and both deterrence and advice are also easier (there is higher certainty of detection and sanction – and clearer

[295] Norms on manufacturing in pre-modern times were, for instance, "specification" norms. Such were also the earliest occupational safety norms, even though we may find that some of the specifications were relatively vague compared to more modern standards.

[296] Pre-1970s, most countries were essentially using detailed specifications. Since then, a number of jurisdictions and regulatory agencies have introduced "target" or "performance" norms (e.g. the US EPA, UK HSE, EU "New Approach" directives etc.). Nonetheless, around the world, the bulk of technical norms tend to remain "specification" ones, e.g. in the post-Soviet space, or in some post-colonial countries (though in this latter category the most frequent problem is the *lack* of technical norms, resulting in excessive enforcement discretion). Many of these specification norms end up being outdated, and/or exhibit contradictions between different regulatory areas or regulators.

recommendations to make). They also offer benefits to firms, particularly smaller ones, which have less resources (both human and financial) to investigate how to be in compliance with "target" or "performance" ones – specific norms can offer certainty and predictability[297]. They tend, however, to create major problems too. First, because of what Baldwin (1995) calls "errors of inclusiveness" – "because they discourage desirable activity (through over-inclusiveness) or they fail to rule out undesirable activity (through under-inclusiveness)" (p. 177). Such situations can arise because of initial rule-design mistakes, but even more frequently occur because rules have not kept up with technical and scientific changes – something which the vast number of rules needed in a "specification" approach makes it likely to happen. This can result in situations where rules directly impose using an inferior technology, not only from a business perspective but from a public welfare perspective – this is the case of a number of Soviet standards that are still in force in many post-Soviet countries, e.g. in hygiene and fire safety, and specifically mandate the use of certain materials, techniques, processes, that were "state of the art" in the 1960s (when the standards were adopted), but have long ceased being so (see International Finance Corporation 2008 on technical regulations in Ukraine, for instance).

Baldwin (*ibid.*) extensively discusses the question of rule-design, and questions the possibility to find an "optimal" degree of rule precision (pp. 176-181). What he also emphasizes is the importance of considering, along with the rule's design and contents, questions of "form, force and type of sanction" as well as "how such problems [of inclusiveness] may be dealt with during the compliance-seeking process" (p. 181). Rightly, Baldwin considers not "rules" in isolation, but rules *along with their enforcement process.* Having discussed the different explanations for over-inclusiveness[298], and the problems involved in addressing them at the rule-making stage, Baldwin suggests that "an alternative response is to write rules that devolve discretion down to enforcers so that issues of inclusiveness are dealt with by selective enforcement", e.g. the famous example of UK health and safety rules based on the notion of "so far as is reasonably practicable" (p. 184). This, however, relies on "high levels of enforcer discretion", raising the twin risks of capture or abuse. It is also far from certain that enforcers will, indeed, be selective, and this "depends on regulatory styles and traditions" (*ibid.*). In addition, it is also essential to consider *which enforcement tools and methods* will be used: "compliance-seekers have at their disposal a number of alternatives to prosecution (e.g. persuading, advising, and promoting) and it cannot be taken for granted that the kind of precise rule that complements a prosecution strategy will be the best kind of rule to use in association with other techniques (p. 178).

Other authors, from different perspectives, fundamentally concur with Ogus (1994) and Baldwin (1995). As Morgan and Yeung (2007) put it: "rules are not self-executing, and scholars have devoted considerable energy to understanding the challenges associated with the use of rules as a mechanism for guiding behaviour" (p. 153). To a large extent, they add, rules are "indeterminate", i.e. their application depends on subjective and contingent factors (*ibid.*). Black (1997), in particular, has written on the ways in which the inherent generalization and abstraction necessary to develop rules results in problems when applying them. Indeed,

---

[297] At least when enforcement is fair and transparent, and there is not a maze of partly contradictory norms. In many post-Soviet countries, for instance, conflicting requirements between e.g. hygiene, occupational safety, construction safety etc. result in situations where economic operators *cannot* be in compliance with all. We observed such situations directly e.g. in Ukraine (conflicting requirements on materials to be used, and on location of garbage disposal, between sanitary and fire inspectors) and in Lithuania (labour inspector demanding that an escape door be locked shut to prevent undue entry, whereas fire safety regulations would mandate that it be free to open in case of evacuation need).

[298] Baldwin lists several possible causes for over-inclusiveness (pp. 182-183, building in particular on Bardach and Kagan 1982): first, "the informational costs of designing rules of optimal inclusiveness are considerable" leading to the tendency to "externalize costs on those who are regulated or on to enforcement officers". Second, a tendency to "risk-regulation reflex" behaviour (see further in this work) that leads to "opt for an across the board solution" in response to "mischief at a particular location". Third, pressure from interest groups. Fourth, the tendency to build on public outrage to a disaster and thus get rules adopted as soon as possible (again, a variation of the "risk-regulation reflex" problem). Fifth, a "regulatory ratchet", through which new rules get added, but old ones are not removed .We would add to these problems of limits of scientific and technical knowledge, regulatory culture (risk aversion), and (as indicated above) the simple effect of time (rules getting outdated).

"the generalization which is the operative basis of the rule inevitably suppresses properties that may subsequently be relevant or includes properties that may in some cases be irrelevant" – and in addition "the causal relationship between the event and the harm/goal is likely to be only an approximate one". Black adopts a squarely instrumental view: "legal rules, and particularly regulatory rules, perform social management and instrumental functions (…) and their success is measured in terms of the extent to which they ensure that the substance of policy is achieved. (…) Under-inclusion can represent 'missed targets'; over-inclusion, excessive intrusion" (pp. 5-15). Overall, there is always an "imperfect correlation between proxy requirements and actual hazards" (Bardach and Kagan 1982, p. 71).

The observation of inspections and enforcement practices suggests that the theoretical impossibility of designing "optimal" rules (that Baldwin 1995 appears to demonstrate[299]) is validated by experience. In fact, the impossibility may be even stronger than suggested in reality, because even very specific and precise norms end up not working uniformly in practice because of differences in enforcement methods. While some agencies and officers will register a violation and impose a sanction even for the slightest variations from the norm[300], regardless of whether it corresponds to a real risk to the public welfare, and of the consequences of the sanctions[301], others will apply a "risk proportionate" enforcement approach.

Thus, there seems to be no escape from enforcement discretion if one is to avoid the twin pitfalls of under- and over-inclusiveness[302]. Ogus (*op. cit.,* pp. 170-171) attempts to find ways to make standard-setting more "optimal", but they all end with relying on regulators and their staff to administer wisely rules written in a more flexible language[303]. Trying to curtail discretion can lead to difficulties in fighting "creative compliance", i.e. formal compliance with specific requirements that "covers" effective undermining of the regulation's objectives (see Baldwin 1995, pp. 185-189). A recent report by the Scientific Council to the Netherlands' Government (*Wetenschappelijke Raad voor het Regeringsbeleid* – WRR) underlined the same risk of "creative compliance" as a major concern, and called in response for *increased* regulatory discretion and less specific norms (WRR 2013).

If only highly damaging (both for the economy and the regulation's own objectives) rigidity can minimize discretion, and if even in such cases discretion can never really be fully avoided, then trying to understand better *how to organize* this discretion is indispensable. This is particularly true considering the very real and considerable pitfalls of unfettered, uncontrolled discretion – regulatory capture on the one hand, abuse and rent-seeking on the other (and corruption and ineffectiveness in both cases)[304].

---

[299] *Ibid.* pp. 179-181, building on Diver (1983).

[300] A case frequently observed directly by the author in post-Soviet countries (see International Finance Corporation reports, various years) – but also often reported by businesses in other countries, e.g. in France with labour inspectors (direct interview with the authors of a recent government review of regulatory inspections – see also Chapelle and Clément 2015).

[301] Taking the above examples: in Ukraine or Tajikistan, many inspectors impose sanctions for the slightest variation from the norms imposing a precise height from the floor for items such as electric sockets or fire extinguishers, even if the variation is less than 1 cm, and has absolutely no risk impact. In France were reported examples of labour inspectors filing a violation and imposing sanctions for every minor discrepancy from the legal work time, regardless of circumstances, significance etc. (and of the impact, which in one case was the withdrawal of a foreign investor from a locally significant business).

[302] As our examples above suggest, there may be no escaping discretion in any case. Even in systems (e.g .US OSHA) which try and minimize regulatory discretion (with a number of side-effects), discretion remains – if not in the hand of regulators, then in the hand of judges called upon to decide conflicts between regulators and businesses.

[303] Ogus (*ibid.*) suggests e.g. the reference to a "general principle" that "may be accompanied by guidelines", or to "confer power on an agency to create formal differentiated standards for invidiual firms or groups of firms" – or to leave differentiation "to the enforcement stage". All of these "solutions" are in fact different ways of establishing and framing discretion (and of assigning it to different organizational levels and operational stages). It remains that it means basically that regulatory discretion is unavoidable if one wants to have at least a "decent" combination of effectiveness and efficiency of rules.

[304] Ogus (*ibid.*) also discusses several of the issues arising around the enforcement approaches and practices. He shows that, in many cases, inspection officials will tend "in exercising their discretion" to "find it difficult to resist arguments for leniency based on grounds" such as financial difficulties, local unemployment etc., even when these are clearly *not* foreseen as factors in the regulation (pp.

A short *coda* is in order to this discussion of why, when and how to regulate: the question of *who* should do it. The importance we have found of regulatory discretion speaks strongly in favour of adequate *professionalism* of inspection and enforcement staff, and maybe also of officials in charge of improving regulatory implementation methods *across the board.* We will discuss the question of inspectors' professionalism in the third part of this research, looking at practical examples. As for the idea of having officials who understand regulatory issues, and are tasked with coordinating and improving them, it can be found e.g. in Breyer (1992). Having shown the importance of expertise and professionalism issues in regulation (see e.g. pp. 49-50) Breyer, looking for solutions towards more effective and efficient "risk regulation", proposes creating a "new career path" of civil servants with expertise in all major regulatory fields, and a "small, centralized administrative group charged with a rationalizing mission" (pp. 59-60). This would, in Breyer's view, help build a more rational, "reformed", "risk-based" mission for regulators (pp. 64-65). We will see in the third part that there have been experiments in this direction, for instance the creation of the UK's Better Regulatory Delivery Office, and that they bear a close relation to attempts to use "risk-based approaches" more systematically.

iii. Conclusion – the importance of implementation – learning from the regulatory field

From the above, we can conclude several points of relevance for the rest of the research. First, regulation has costs as well as benefits, and it has limitations – and this applies also to various types of regulation or regulatory instruments. Within this framework, "optimizing" regulation's effectiveness and efficiency appears to require leaving a significant space for discretion in enforcement – thus *how to structure* this discretion is an important question. This is what we will focus on in the rest of this research.

Second, regulation can legitimately be treated as a *specific* field, and there are sound reasons to apply strict scrutiny to limitations of economic freedom and their potential adverse effects, but not more so than would be true for a number of other legislative fields. Rather, the specificity of regulation has emerged to a large extent as a result of the interplay of conflicting actors, and the salience of economic and social issues it relates too. There may thus not be real legitimacy to treat regulatory issues *differently* than we would other fields of legislation that impinge on fundamental rights or freedoms – but maybe there are ideas and lessons that have been developed in the study of regulation that have emerged there more strongly (precisely because of conflicting interests, salience of issues etc.), and that could be applicable to other fields. We will see when discussing compliance theories that much can be learned from non-regulatory fields (e.g. interactions between citizens and the police). It may well be that much could in turn be learned from "better regulation", "smarter regulation" and "risk-based inspections" that could be applicable to interactions between the state and citizens, civil society organizations, the media or other stakeholders, in a variety of fields. This could be the case of the requirement to analyse costs and benefits, demand extended consultations and discussions before

211.212 – note that the 2014 UK Regulators Code now explicitly *mandates* that all regulators should have regard to economic impacts when taking their decisions – see: https://www.gov.uk/government/publications/regulators-code). We have seen, on the other hand, many countries where regulatory officials did not pay *any* attention to such issues in most cases (see above Ukraine, France etc. examples). Ogus also mentions how command and control regimes correspond to "power, prestige and job satisfaction" for regulators (p. 256), something which is even more true in countries where corruption issues are significant. Again, these different (somewhat conflicting) pitfalls all make it only more important to study more closely how regulatory discretion can be better understood, "framed" and managed.

issuing new rules, focusing on key risks and leaving more room for voluntary compliance when risks are low, modulating enforcement responses based on risks etc.

As Voermans (2015) has shown, the lack of consideration of issues such as compliance drivers and methods to assess, understand and improve compliance levels is one of the roots of implementation problems for European legislation (see pp. 357-359 in particular). This is an interesting case where considering the best practices in regulation and regulatory enforcement could greatly benefit a broader field of legislation, and a "higher" institutional level (since EU legislation applies primarily to Member States, and not only or firstly to individuals). Thus, the lessons learned in studying regulatory practices, and in particular regulatory enforcement and inspections, could be found to have broader relevance – to other legislative and policy fields, and to a variety of actors and institutional levels.

Finally, a short point is in order to clarify the exact place of *inspections and enforcement* within the context of justifications for regulation. There are several ways in which inspections and enforcement are relevant when considering the need for and legitimacy of regulation, and in deciding on the most appropriate method. First, when the potential costs of a given regulation are assessed (both for the state, the duty holders, and the economy at large), it is essential to consider the specific costs of the "enforcement" stage, and for this to decide between different inspections and enforcement approaches (including the "none" option)[305]. Second, when the expected effectiveness of a given regulation is envisioned, the inspections and enforcement stage is equally important (and, again, there are several options with different expected results). Third, inspections and enforcement can also be considered *independently*, when a regulatory framework already is in place and there is no discussion of its being revised. In such circumstances, different inspections and enforcement choices (in terms of institutions, resources, approaches, "on the ground" methods etc.) will present very different costs, expected outcomes, and also levels of restrictiveness and intrusiveness, and thus can be subject to the same kind of analysis as would be done of regulations themselves in terms of both legitimacy and adequacy.

b. Costs and effectiveness of regulation and enforcement – theory and evidence

In a number of ways, costs and effectiveness questions are central to regulatory discussions, and regulatory inspections and enforcement issues are no exception. We have exposed in the first section the ways in which regulatory inspections were explicitly set up in order to *address perceived problems,* meaning that, given this utilitarian purpose, *effectiveness* is a central consideration. Furthermore, as we will highlight in the third section (covering experiences of risk-based inspections and reforms), claims to *reduce costs* (to private businesses, citizens, the economy at large etc.) of inspections (and, by extension, of regulation) are central to the drive for more risk-based inspections. At the same time, *effectiveness* problems are also important to the risk-based inspections discussion, and have given the impetus to many changes in rules and practices. There are only few research undertakings that have focused specifically on assessing *inspections' effectiveness*, and even fewer that have looked seriously at costs (a topic on which, on the other hand, there is a certain amount of publications from governmental and inter-governmental institutions). We will consider these in the third section, but for now we will briefly review the broader accounts of *regulation*, under which inspections and enforcement are generally subsumed, to see what they can tell us about these issues.

In fact, assessing effects (positive or negative) of regulation is easier said than done. Even though regulations have been given far more prominence in public discussions and research over the past couple of decades,

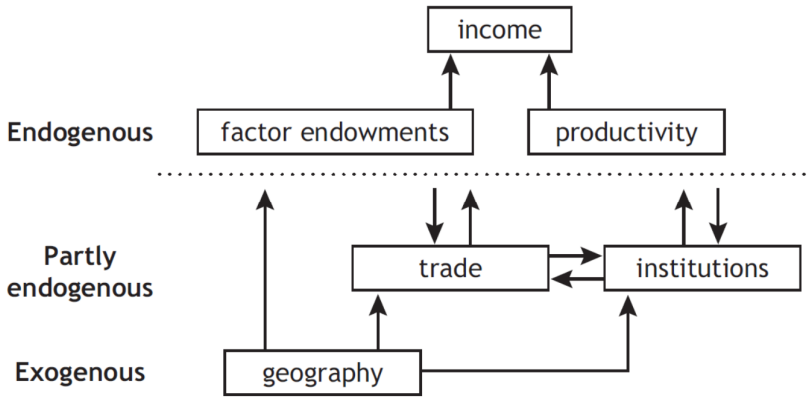
[305] See OECD 2014 (b), in particular the principles on "evidence-based enforcement" and "selectivity"

there is little solid proof or undisputed evidence that they make a considerable difference – be it to economic growth, or to the public welfare they aim at supporting. While this may sound provocative or even contrarian, and the parallel progress of regulation and welfare over the past couple centuries may seem to be proof enough, there are in fact a number of studies that cast doubt on causal relationships around regulation. If regulations themselves are of limited relevance, then one may argue enforcement and inspections are also matters of secondary importance. We cannot pretend to make a comprehensive review of the literature on these topics, but will try and cover them briefly to show some of the main findings and problems, and look at reasons the issue may matter regardless of the "inconclusiveness" of economic studies.

Regulatory reform and growth – context and specific and content of reforms matter

First, the impact of regulations on economic growth, competitiveness or jobs, is indeed disputed. On the one hand, Djankov, McLiesh and Ramalho (2006) write that "our results indicate that government regulation of business is an important determinant of growth" and that "relationship between more business-friendly regulations and higher growth rates is consistently significant in various specifications of standard growth models, and more consistently so than other determinants commonly used in the growth literature". They conclude that "Our results also have significant implications for policy [and] suggest that countries should put priority on reforming their business regulations when designing growth policies". Many other economists, however, beg to differ – both with the findings about the significance of regulatory issues, and about the policy prescriptions pushed by Djankov *et al.* (which all link to the *Doing Business* report, a project which Djankov long headed).

From a general perspective, there is consensus that "institutions" (part of which are regulations, and regulatory inspections and enforcement) are *one of the components* of growth. There obviously are many other components, including geography, demography, social and cultural factors, technology – but institutions have an important role, and interact with many other factors. Rodrik (2003) shows the ways in which good institutions supports growth in several ways (by directly impacting productivity, and through its effects on trade) in the following chart (*Introduction,* p. 5 – figure 1.3).



Rodrik, looking at the case studies gathered across the world, writes further that "institutions that provide dependable property rights, manage conflict, maintain law and order, and align economic incentives with social costs and benefits are the foundation of long-term growth. This is the clearest message that comes across from the individual cases" (*ibid.*, p. 10). However, there is far less of a "standard prescription" of what exactly these "institutions" should entail than there is in Djankov *et.al.* Indeed, as Rodrik indicates further, "good institutions can be acquired, but doing so often requires experimentation, willingness to depart from orthodoxy, and attention to local conditions (…) Perhaps nowhere has this been clearer than in China. Qian's

discussion of China focuses on what he calls "transitional institutions"—institutions that can differ greatly from off-the-shelf, "best practice" institutions (…) [and] can have the virtue of being more suited to the realities on the ground on both economic efficiency and political feasibility grounds. Qian shows that the Chinese leadership experimented and purposefully crafted imperfect, but feasible institutional arrangements (…) [which] succeeded because of their high ratio of economic benefits to political costs" (*ibid.,* p. 13).

A particular area of focus for reform is that of product-market regulations, and this is an important one for our research, as many inspections relate to product-market rules (e.g. food safety, non-food products market surveillance etc.). The findings are generally quite consistent that improvements in this area have a strong positive impact on productivity. These productivity-boosting effects of making product-market regulations more flexible are visible not only in the sectors directly affected, but "downstream", i.e. liberalizing production/intermediary goods has effects on the productivity of all the sectors that use these outputs: "Regulations that bridle access to otherwise competitive markets and unnecessarily constrain business operation can be a drag on productivity growth. While most analyses of this issue have focused on the effects of these regulations on the productivity of the firms or sectors directly concerned, the main point of this paper is that such regulations can also have powerful indirect depressing effects on the productivity of other sectors through input-output linkages" (Bourlès, Cette, Lopez, Mairesse, Nicoletti 2010, p. 28).

In short, regulatory issues (and particularly product-market regulations, which have a strong link with inspections) are relevant to long-term growth prospects, among a number of other drivers – but exactly in what way, and what improvements are most important, is likely to depend significantly on the broader country context. What, then, of the importance of such issues to *developed* countries, and to economic recovery in the current crisis (one dare not yet say "post-crisis") context? While the European Commission (EC) (EC 2014) and the OECD (OECD 2015 a) frequently emphasize the importance of "structural reforms" (which include employment law, tax administration, product-market regulations etc. – and which as a result cover reforms in a number of inspections), the prominence of discussions of these structural reforms in a "crisis recovery" context is somewhat misleading. Even the EC and OECD take pains to remind readers that these reforms are not "quick fixes": "Structural reforms to labour and product markets help to improve economic growth prospects and the ability of economies to adjust to shocks by expanding flexibility and improving the efficiency of how and where productive factors are used. The recent financial and economic crisis prompted EU countries to under-take considerable reforms, which are now starting to show tentative results. Their full benefits, however, may take years to materialise, which means that governments must avoid the temptation to give up on them now that the economic situation is somewhat more comfortable" (EC 2014, p. 1). The OECD states that: "overall, structural reforms implemented since the early 2000s have contributed to raising the level of potential gross domestic product (GDP) per capita by around 5%, with most of the gains coming from higher productivity" and that "further reform (…) could further raise potential GDP per capita by up to 10% on average across OECD countries" (OECD 2015 a, p. 106).

The IMF take is somewhat similar, but more precise and grounded in more economic analysis. In its latest take on the issue (IMF 2015), it indicates: "The analysis illustrates that structural reforms in the euro area can increase its real GDP markedly, though it may take time for their full potential to be achieved. Structural reforms are critical to improving the long-term capacity of economies to grow through both more intensive use of resources and higher productivity". Within these reforms, the "largest gains for euro area countries could come from product market reforms" (where inspections are an important aspect). However, the IMF cautions that "Weak demand conditions may dampen the already small short-term impact" (IMF 2015, Ch. 7, p. 22). Some independent analysts[306] are far more critical of the idea that structural reforms are what is

[306] As EC, OECD and IMF have all been advocating structural reforms, there is an incentive for their publications to be moderate in their skepticism of such reforms' impacts.

urgently needed in a time of recession. They note that "a broad consensus has emerged: Peripheral euro-area countries need to urgently adopt structural reforms that increase competition in product and labor markets" (Eggertsson, Ferrero and Raffo 2013, p. 2). Their conclusions, however, show that, while "structural reforms can greatly reduce the competitiveness gap between the EMU core and periphery and boost income prospects in the region", "the timing of such reforms is crucial. If undertaken during a crisis that takes monetary policy rates to the ZLB[307], structural reforms can deepen the recession by worsening deflation and increasing real rates" (*ibid.,* p. 32). In summary, "in a crisis that pushes the nominal interest rate to its lower bound, these reforms do not support economic activity in the short run, and may well be contractionary" (*ibid.,* p. 1). While the authors consider all kinds of structural reforms (product markets and employment) together, and there may well be differences between the effect of these two types (with most of the demand depressing effects coming from employment reform), caution remains in order.

Thus, to summarize these overall findings about the positive impact of regulatory reform on growth: it is likely to be significant (possibly major) in the long term, however its impact on the short term is more limited (and particular caution is needed in times of recession due to a shortfall in demand), and the exact contents of the regulatory reforms that will be effective is highly context- and country-specific.

Regulations impact on competitiveness, growth and jobs – more complex than it may seem

A related contention to the one that regulatory reform is "good for growth" is that "regulations" (or at least their abuse) would be "bad for competitiveness". While this appears to be grounded in simple logic (if you add more hurdles and demands on businesses, their costs of operating should be higher, and this in turn should make them less competitive globally), findings again show that this is not as clear-cut as it may seem.

A first problem is that, while regulations impose costs, it is not clear how high they are. Even the Australian *Taskforce on Reducing Regulatory Burdens on Business*, which had a clear interest in showing the relevance of its own task, had to acknowledge that "while a number of studies have sought to estimate the economic costs of regulation in Australia, the limitations of such studies mean that the estimates should be treated with caution" (Banks 2006, p. 12). This same report, however, suggested that "the economic cost of complying with regulations is a key determinant of national competitiveness and the investment environment for businesses. These costs can be direct, such as capital and operating costs. They can also be indirect, that is, opportunity costs, where the principal(s) of the businesses are taken away from their strategic roles of driving innovation, securing investment and increasing productivity" (*ibid.,* p. 11). Thus, the importance of improving regulations is predicated on their (negative) impact on innovation, investment and productivity – and, through these, competitiveness.

It is, then, worth considering the literature on one of the most prominent areas of regulation, environmental rules, and their impact on competitiveness. A quick review of findings gives a picture that contrast sharply with the above emphasis on regulations as a serious problem: environmental regulations appear to have very limited, if any, negative impact – and some studies even suggest a positive long-term impact. We summarize a few interesting studies, not in the aim of reaching strong conclusions, but in order to show the complexity of the topic, and the many factors that may influence findings. From our perspective, these apparent contradictions are interesting because they may point to the importance of (too often neglected) *implementation* questions in order to understand the impact of regulations.

The finding that environmental regulations have (at worst) little negative impact on competitiveness is relatively constant through repeated studies over a decade. One of the earliest studies (Jaffe, Peterson,

[307] Zero Lower Bound

Portney and Stavins 1995) concluded that "there is relatively little evidence to support the hypothesis that environmental regulations had had a large adverse effect on competitiveness, however that elusive term is defined" (p. 157). The conclusion is not entirely one-sided, though, and they add that "long-run social costs (…) may be significant, including adverse effects on productivity" – but studies looking at "exports, overall trade flows and plant-location decisions" show impacts that are "either small, statistically insignificant, or not robust to tests of model specification" (*ibid.,* p. 158). The authors, however, have interesting insights on *why* this may be so – i.e. why indeed these impacts may be small in reality, and why there may be measurement issues. We will come back to these a bit later.

More recent reviews report roughly similar findings. A review prepared for the United Kingdom's ministry in charge of the Environment (Department for the Environment, Food and Rural Affairs – DEFRA) in 2006 provides with nuanced, interesting points (SQW Limited 2006). On trade, it writes that "the evidence seems consistently to be that the costs imposed by tighter pollution regulation are not a major determinant of trade patterns even for those sectors most likely to be affected by such regulation. However, there is some evidence that regulatory stringency may exercise an influence once account has been taken of the factor intensity of the different industries and the relative factor abundance of countries. Thus, for a country in which a specific production factor is relatively scarce and an industry intensively uses this factor, then even a modestly stringent environmental regulation will induce a decline in exports" (p. iii). At the firm-specific level, the report finds that "there is a modest productivity penalty in the short term associated with increased stringency of regulation. But, they also provide evidence of a countervailing innovation push over the longer term – especially in larger firms with a track record of innovation" (*ibid.,* p. iii). In macro-economic terms, the conclusion is that "regulations are unlikely to increase competitiveness (…) and may adversely affect it" but "the adverse effect can, to varying degrees, be offset" – through tax incentives, multilateral agreements with "competing nations and regions" or (more interestingly for our research) by ensuring that "businesses are made aware of the regulations" and prompting "them (through advisory and grant support) to invest in improved operating practices" (*ibid.* p. ii).

A more recent research paper supports further the same views (Dechezleprêtre and Sato, 2014). The paper states that "environmental regulations can reduce employment and productivity by small amounts, in particular in pollution- and energy-intensive sectors, at least during the transitory period when the economy moves away from polluting activities and towards cleaner production processes. Job effects are more likely to occur within countries, where relocation barriers are low, than across borders" and that "over the longer run, when macroeconomic adjustments, geographical and sectoral reallocation are factored in, job effects are even smaller than in the short run" (p. 3). The authors add that " There is little evidence to suggest that strengthening environmental regulations deteriorates international competitiveness. The effect of current environmental regulations on where trade and investment take place has been shown to be negligible compared to other factors such as market conditions and the quality of the local workforce. However, the impact could increase in the future if efforts to control pollution diverge significantly across countries" (*ibid.*). The authors go on to add that "benefits of environmental regulations often vastly outweigh the costs".

A last research paper is worth quoting on the economic impact of environmental rules. In this paper (Bivens 2012), the author argues that "when significant economic slack persists even when the interest rates controlled by the Federal Reserve are held at zero, the overall effect of cost-raising regulatory changes is almost surely expansionary" (p. 2). In other words this suggests that (at least from a macro-economic modelling perspective), when the economy is in recession (producing below its output potential), i.e. suffering from a lack of demand, raising regulatory demands can act as a kind of "stimulus", because the need to invest in order to comply with the new rules would generate a form of additional demand. Since the economy is not capacity-constrained in such a setting, but demand-constrained, the new rules can have a positive economic impact (rather than a slight negative). As a result, "the effects of some specific regulatory changes (…) are

surely positive for job creation". This is in line with what has been called the "Porter hypothesis", formulated by Michael Porter and co-authors in a number of important papers (see e.g. Porter and van der Linde 1995, Esty and Porter 2005). Quoting Esty and Porter (2005, p. 425), "our findings suggest that the environment need not be sacrificed on the road to economic progress. Quite to the contrary, the countries that have the most aggressive environmental policy regimes also seem to be the most competitive and economically successful. Moreover, we find preliminary evidence that countries that adopt a stringent environmental regime relative to their income may speed up economic growth rather than retard it."

Making sense of apparent contradictions

This short "review of reviews" leaves us with what can be a counter-intuitive result for many: regulations (in this case, environmental, but there is reason to think that the same mechanisms may apply to many other types of regulation) seem to have at worst a very limited impact on competitiveness, trade and macro-economic results – and may even have in some circumstances (persistent economic slump) a positive impact on jobs and for some firms (and with the right type of context, support etc.) a positive impact on adoption of latest technologies and thus competitiveness. It may seem somewhat conflicting with the previous findings, which suggested that regulatory reform (product-market regulations mostly) would have a positive impact on growth and competitiveness.

The solutions to this apparent puzzle may lie in at least three directions: differences in the nature of industries affected, flaws in studies and data and a difference between the "level of regulation" overall (i.e. the substantive requirements embodied in the rules) and the specifics of regulation i.e. how it is worded, which instruments are used to implement it, and how control and enforcement are handled (and support provided, or not). The second and third point, in particular, are highly significant for our research.

First, for some industries, the cost of compliance may be far higher than for others – and/or international competitive pressures may be higher. As Jaffe, Peterson, Portney and Stavins (1995) put it "for all but the most heavily regulated industries, the cost of complying with federal environmental regulation is a relatively small fraction of total cost of production" (p. 158) – but this means that for *some* industries the effects may be far stronger. The same authors add that "although U.S. environmental laws and regulations are generally the most stringent in the world[308], the difference between U.S. requirements and those in other western industrial democracies is not great" – and that "even where there are substantial differences" in environmental requirements, U.S. and multinational firms "are reluctant to build less-than-state-of-the-art plants in foreign countries", at least after the Bhopal disaster. They further contend that "even in developing countries where environmental standards (and certainly enforcement capabilities) are relatively weak, plants built by indigenous firms typically embody more pollution control (…) than is required" (*ibid.*). These last points raise several concerns, since it is far from clear that these different points are all true across the globe now, if they ever were. Certainly, the major changes in the global economy in the past couple decades, and in particular the relocation of a substantial part of manufacturing activities to China and other countries combined with persistant reports of "less-than-optimal" compliance with a number of safety and environmental standards in these countries, suggest that the findings of some studies may not hold true anymore. Given that many reviews of evidence incorporate studies that are years or sometimes decades old, this may weaken their findings.

Second, data limitations are significant, and may explain the variations in findings (and in other words mean that many "findings" are no such things, but rather illusions caused by inadequate data). Quoting once more from Jaffe *et al.* (1995): "in many of the studies, differences in environmental regulation were measured by

[308] This may not always be the case anymore as in a number of cases EU regulations for instance are more stringent.

environmental control costs as a percentage of value-added, or some other measure that depends critically on accurate measurement of environmental spending. Even for the United States (…) compliance expenditure data are notoriously unreliable. The problem is even more pronounced in other OECD countries (…) Thus, we may have found little relationship between environmental regulations and competitiveness simply because the data are of poor quality" (p. 158). As we have purported to show on the section on practical data limitations, the quality of much of the data on regulatory issues, compliance burdens etc. is of very poor quality, be it due to the difficulty or unwillingness of respondents to answer correctly, or because of lack of quality control etc. In any case, it seems that we may face another case of the tendency of many studies to draw major conclusions from calculations based on faulty data. Looking at the broader picture of both quantitative and qualitative findings, and at longer time periods, international comparisons of "growth trajectories" etc. may thus be the best we can do at this stage to compensate for these data problems.

Finally, and this is the most significant for this research, the (apparent) contradictions in findings may be to a significant extent due to the lack of attention of most studies to the major distinction between the underlying technical requirements (what businesses are supposed to achieve, substantially) and the procedures, processes and regulatory instruments associated with these requirements (what permits and licenses have to be obtained, through which processes, how controls are conducted, what avenues exist for redress etc.). Quoting Jaffe *et al.* (1995) one last time: "only two of the studies we reviewed controlled for differences in "regulatory climate" between jurisdictions. If the delays and litigation surrounding regulation are the greatest impediments (…) these effects will not be picked up by studies that look exclusively at (…) standards or (…) spending" (p. 158). The DEFRA study (SQW Limited 2006) likewise noted "that there is not a great deal by way of empirical work on the different forms that regulation can take and the effects of their form of implementation on firm behaviour" (p. 41). Further, it adds that "Regulation design, stringency and efficiency can influence the relationship between environmental regulation and competitiveness. Stringency may well be less important than the design of regulation itself" (p. 46-47). The same study also suggests that "awareness" of regulations may play a role in mitigating possible adverse effects (p. 37), but does not go in any further details.

What these points all suggest is that existing studies may have focused far too much on either the underlying requirements themselves ("standards"), regardless of their enforcement context, or on the estimated costs (with the associated problems in data quality). It may thus well be that researchers have been ignoring one major direction of inquiry: how regulations are "delivered", and what effects this has – and in particular, what role inspection methods play in this. We will see in the next section that such attention to regulatory instruments appears to be very fruitful in the case of licensing.

The specific effects of regulatory instruments – the example of licensing

Looking at the effects of "regulation" *in general* is, in our view, inherently problematic, since it assumes that the ways in which regulations are implemented are largely indifferent (and it also assumes, more or less, that regulation *is indeed implemented* – which can be a heroic assumption indeed). Rather, there is evidence that *specific regulatory instruments* can have different effects, both positive and negative – but there is still only limited research on this, and on the comparative costs and benefits of different instruments.

From an economic perspective, the most significant research has been conducted on licensing, mainly by Morris Kleiner[309] and under his direction – and this research has recently gained a higher public profile, and

[309] See in particular Kleiner and Kudrle 2000, Kleiner 2006, Kleiner and Krueger 2010, Bryson and Kleiner 2010, Kleiner and Krueger 2013.

been taken up and summarized in a report of the US Council of Economic Advisers[310]. This research is primarily done from an economic perspective, and considers what the costs of licensing are in terms of employment effects, and its effects in terms of increased health, safety etc. Kleiner *et al.* look at what is called in the US "occupational licensing", i.e. the licensing of professional occupations (generally licenses given to individuals), and not to the entire range of license types – of buildings, economic operators[311] etc. Kleiner *et al.* do not question the scope of licensing by looking at the actual level of risks (even though this is one possible approach), but rather use economic models to capture the comparative effects of licensing and other, "milder" regulatory instruments, such as certification of practitioners (which is similar to licensing, but voluntary – i.e. consumers can choose between certified and non-certified practitioners), or registration (which is a significantly weaker instrument, only resulting in a catalogue of practitioners, but not necessarily indicating competence).

To compare these, Kleiner *et al.* look at (a) comparable types of activities, some of which are licensed, and the others not, in the same jurisdiction – (b) identical activities in different jurisdictions, some of which require a license, and others not – (c) identical activities in the same jurisdictions before and after a change in legislation which altered the licensing regime. The studies then compare different outcomes that can be linked to licensing (or its alternatives): effectiveness in terms of achieving social welfare goals on the one hand, and employment and income effects. If licensing is seen to produce significantly better outcomes, e.g. better dental health where dentists are licensed (or where licensing requirements are more strict), it may balance the costs that it imposes. Conversely, if the economic effects (reduced employment, higher "rent" for licensed professions) are very significant, and the benefits marginal, this may lead to questioning the appropriateness of licensing.

The potential (and purported) positive effects of licensing stem both from its direct "screening" effect (expected to improve quality, health and safety) and from its market information effects (increasing consumer trust and thus potentially increasing consumer demand for specific goods and services where it may otherwise remain low because of information asymmetry)[312]. The two aspects are complementary. The former (safety) that is generally put forward as the main justification for introducing (or tightening) licensing requirements.

---

[310] See Department of Treasury, Council of Economic Advisers and Department of Labor (2015), available at: https://www.whitehouse.gov/sites/default/files/docs/licensing_report_final_nonembargo.pdf.

[311] "Licenses" or "permits" (two terms which come from different latin words both meaning that something is allowed) are regulatory instruments that are used for a number of issues, in many different ways and for a variety of purposes. One can at least distinguish licenses applying to buildings, premises and equipment (construction permits, licenses/permits for specific machinery, facilities etc.), licenses applying to operators (banking, tourism, television, mobile phone operators etc.), and licenses applying to individuals (doctors, hairdressers, taxis etc.). Licenses may be open-ended or time-limited, they may be issued in unlimited numbers or submitted to a *numerus clausus*, they may or may not require a number of documents, qualifications, fees etc. Their goals may combine safety and protection against risks with economic objectives (managing scarce resources), etc. Finally, some forms of entry regulation can be understood by some (e.g. Kleiner) as equivalent to licensing, even though no actual license is issued. This is the case of what many EU countries call "regulated professions", whereby the exercise of some professions does *not* require a specific license to be issued *but* requires some qualifications including e.g. a state-sanctioned diploma, a certain number of years of exercise etc. For clarity, we would advocate to distinguish such regulated professions from licensed occupations – both are restrictive regulatory instruments, but somewhat different. The same goes for self-regulated professions, e.g. doctors or lawyers in a number of EU countries, which are not officially called "licensed", even though the effect is similar.

[312] Kleiner 2006: "existence of licenses may minimize consumer uncertainty over the quality of the licensed service and increase the overall demand for the service" (p. 1). Department of Treasury, Council of Economic Advisers and Department of Labor 2015: "Even when health and safety are not an issue, increasing consumer information through regulation can be beneficial. If consumers are unable to distinguish between high- and low-quality providers before purchasing a good or receiving a service, low-quality providers can remain in the market without being recognized as such, reducing the average quality in the market and reducing the incentives for other providers to invest in quality improvements.12 Furthermore, if consumers are sufficiently concerned about getting a low-quality provider, then informational uncertainty may depress demand for goods and services. Consumers who would otherwise purchase a product if they knew it were high-quality might forgo their purchase if the quality were uncertain. Licensing is one possible way to address these problems through forcing providers to meet certain quality benchmarks, and creating greater incentives to invest in increased training and skill development" (p. 11).

Arguments in favour of licensing invariably put forward risks and safety justifications – licenses being supposedly necessary to avoid a number of catastrophes (see Kleiner 2006 p. 1). In the original emergence and spread of the licensing system in the US, consumer demand for information on the quality (safety) of service for critical professions appears to have played an important role (*ibid.*, pp. 22-23). However, the resulting "licensing map" does not necessarily suggest a strong match between licensing requirements and hazards: to take a typical example, a taxi driver needs one, but not the manager of a chemical plant, even though the latter most likely has a stronger potential "risk level". Consumer information and addressing information asymmetries may appear to be a better match with actual licensing practices (a company has time and means to screen applicants when hiring a manager, not so with someone hailing a cab on the street, or with a patient urgently looking for a doctor). Remains to be seen whether the beneficial effects of licensing are actually observed *in practice.*

Kleiner, in 2006, used one example of licensed occupation that, with hindsight, shows perfectly the limitations of licensing's effectiveness as a regulatory instrument. Quoting work by Wheelan on occupational licensing in Illinois, Kleiner pointed out, as a good example of "capture" of licensing, the parallel rise in the secondary mortgage market and increasing level of regulation on mortgage brokers (*ibid.*, pp. 46-47). Given what we have seen in the meantime, it is clear (with hindsight) that mortgage brokers licensing did essentially nothing to ensure the adherence to strict standards of practices. A more systematic look at the evidence likewise suggests that regulatory capture is a stronger predictor of actual licensing patterns than public interest.

Quoting the Department of Treasury, Council of Economic Advisers and Department of Labor's report, "with the caveats that the literature focuses on specific examples and that quality is difficult to measure, most research does not find that licensing improves quality or public health and safety" (Department of Treasury, Council of Economic Advisers and Department of Labor 2015 p. 13). Similarly, Kleiner (2006) had already found that "the analysis of studies of licensed occupations finds that the impact of regulation on the quality of service received by consumers is murky, with most of the studies showing no effects on average consumer well-being relative to little or no regulation" (Kleiner 2006 p. 63). Crucially, in spite of being (as acknowledged by the Department of Treasury, Council of Economic Advisers and Department of Labor's report) fragmentary and partial, studies have focused in a number of cases on examples where strong public safety and health effects were claimed (e.g. dentistry), and found them to be at most very limited, and often wholly lacking. Economic impacts, however, tend to be strong – and to support a "capture" view of licensing, with economic welfare for the whole of the population (or country) decreased, but rents for the licensed professionals increased. Indeed, "there is compelling evidence that licensing raises prices for consumers" (Department of Treasury, Council of Economic Advisers and Department of Labor 2015, p. 14) – and "monopoly power [of licensed occupations] may reallocate income from lower-income customers to higher-income practitioners" (Kleiner 2006 p. 59), meaning that licensing has in many cases a negative distributional impact (increasing inequality). In addition to price and income distribution effects, "licensing affects who takes what job. If licensing places too many restrictions on this allocation of workers, it can reduce the overall efficiency of the labor market. When workers cannot enter jobs that make the best use of their skills, this hampers growth and may even lessen innovation" (Department of Treasury, Council of Economic Advisers and Department of Labor 2015, p. 12). Considering the evidence on employment effects, Kleiner (2006) finds that "within an occupation, the employment growth rate is approximately 20 percent higher in states that do not require licensing, but impacts differ widely based on the methods and occupations" (p. 149). On balance, there appears to be a substantial redistribution effect from the general population to the licensed occupations (estimated by Kleiner at $116-139 billion – *ibid.*) and significant lost output due to misallocation of resources (estimated at $34.8-41.7 billion – *ibid.*).

Thus, while licensing appears to have at best limited positive impacts on public safety and market trust, it has clearly demonstrated negative impacts on income distribution and on resource allocation. What is important

is that these are not effects of what is too often, and indiscriminately, called "regulation" – but of *a specific type of regulatory instrument*. The choice of regulatory instruments, and of their characteristics, is thus important, independently of the content of substantive regulations that economic operators have to abide by in terms of practices, safety etc. Interestingly, the Department of Treasury, Council of Economic Advisers and Department of Labor 2015 report contrasts licensing with inspections – noting that licensing board tend to conduct only limited oversight of the license holders after issuance, but that inspections can constitute an effective (and possibly less burdensome and economically harmful) regulatory instrument in place of licensing (pp. 43-44). This is an argument that, as a reform practitioner, we have seen discussed in a large number of countries, e.g. in Greece and Ukraine in recent years[313]. While having merit at first glance (periodic inspections are more likely to help in verifying and supporting sustained compliance than a licensing check administered before start up, and with permanent validity), this leaves aside the question of whether inspections are really effective, and how to make them more so. This shows once again the importance of conducting more research on the specifics of regulatory instruments, and not only on "regulation" considered as an indistinct block.

Regulations, inspections and corruption

In developing countries and emerging markets in particular, but also in high-income economies, regulations can, in a number of instances, be associated with corruption. Inspections, being one of the main points of contact between regulators and regulated businesses, are often associated with corruption in regulatory dealings. While we cannot do justice do this important topic within this research, of which it is not the focus, we will attempt to indicate a few of the ways in which the link with corruption makes the improvement of inspection practices particularly relevant. First, however, let us give a somewhat more precise meaning to the highly loaded and polysemic term of "corruption".

In its broadest sense, corruption can be understood as any way in which the regulatory, legal or administrative process is made to serve a purpose that is fully different from its stated aims, and to function in a way that is in contradiction with its official rules. In a somewhat narrower meaning, which is the one of interest here (and the most commonly accepted one), corruption is when a process or rule is subverted in order to serve specific private interests, for private gain (financial mostly, but possibly political etc.). Money does not always need to change hands, and corruption in inspections certainly does not always mean that bribes are given during the inspection visit. Corrupt behaviours can involve gifts, employment, expectations of future "tit for tat", or any variety of favours, from the business side. From the regulator's side, they can involve turning a blind eye on violation, interpreting rules leniently, harassing competitors, or simply doing one's job normally (in cases where regulators abuse their powers systematically against those that refuse to "pay up" or "play the game")[314].

Regulations and regulatory processes are not the only *locus* of corruption, of course, and poorly structured rules and institutions are definitely not the only (or the main) cause of corrupt behaviours. They are, however, one of the most important areas of corruption (along with police interactions), because both of the large number of rules and regulatory instruments, and because they affect economic activity, and thus present strong opportunities for rent-seeking behaviours (both for regulators and regulated entities). Corruption in inspections generally presents important differences with, for instance, corruption in rule-making. The latter

[313] There is a lot of evidence that the problem of the right use and design of licensing, and more broadly of regulatory instruments, is essential to developing countries and emerging markets. There is rather little academic literature on this topic, and not always recent (see e.g. Ogus and Zhang 2005, Zhang 2009). Practical reform work done by the World Bank Group has highlighted repeatedly the importance of the topic, but literature produced is mostly focused on "how to reform" rather than on an analysis of the *pro* and *contra*. See nonetheless World Bank Group 2006 and

[314] See Ogus (2004), *Introduction* and section *Definition and Typology of Corruption*, for a discussion of the different meanings and types of "corruption".

usually involves high-level capture by major firms, aimed at keeping competition out and/or building captive markets, and (at least in middle- and high-income countries) is likely to involve more "revolving door" offers for officials than outright bribes. Inspections, by contrast, offer opportunities for "decentralized" corruption, involving front-line inspectors, small and medium firms, a variety of "gifts" and favours. In some countries, corruption is essentially the default setting: inspectors go from firm to firm, bribes or gifts are expected, and in their absence enforcement will be ruthless, if needed "making up" violations where there are none.

Unfortunately, the topic is notoriously difficult to investigate, since reliable data on corruption is, nearly by definition, hard to come by. Survey data, for instance, can be very misleading in countries where businesses have reasons to believe that being open about corrupt behaviours could end up creating problems for them (which means the majority of countries). That said, some evidence exists, as gathered for instance in successive surveys by the International Finance Corporation in post-Soviet countries[315]. Even though inspections-related corruption, much as petty corruption more generally, is primarily a problem for developing countries, some high-profile scandals should warn against complacency in high-income, developed countries – for instance the crane inspections scandal in New York City[316]. It is also worth remembering that corruption can manifest itself in misuse of administrative power not for private gain, but for the "profit" of the institutions themselves, as has been abundantly demonstrated in recent years in the United States by the accumulation of fines intended not to deter crime but to fill municipal coffers, and by the abuse of the "civil forfeiture" programs to the "quasi private" benefit of local police departments[317].

Clearly, corruption is linked to a multiplicity of factors: prevailing cultural norms, income levels and distribution patterns, strength or weakness of institutions, social structures etc. It remains nonetheless that rules and regulations, as well as regulatory practices, also have their importance in creating or sustaining corrupt behaviours. Simply put, if rules are impossible to comply with because they are obsolete, excessively demanding considering available resources, overly complex and prescriptive, or any combination thereof, corruption will be the way through which the economy manages to somehow function *in spite* of the rules (much as smuggling is the "natural" consequences of duties that exceed an "economically optimal" level, and smuggling thus rises when duties go over a certain point). Similarly, procedures that are excessively long, opaque, burdensome, and leave too much unchecked, arbitrary power to regulatory officials will tend to lead to abuses of power and corruption, with regulators tempted by rent-seeking, and businesses seeing it as the easier (or the only) way out[318].

Little research exists at this stage on corruption specifically in the context of inspections and enforcement, and as we pointed out already data is often unreliable and makes this a difficult topic to investigate with precision. There is, nonetheless, a body of work on regulation and corruption, that shows the relevance of the issue. Djankov *et al.* (2001) have shown, in particular, how excessive business entry regulation, disconnected from a clear purpose in terms of social welfare, can result in increased corruption and serious economic harm. They write that "in principle, the collection of bribes in exchange for release from regulation can be efficient [from

---

[315] See successive surveys from 2003 onwards in Tajikistan, Ukraine, Kyrgyzstan in particular.

[316] See the official account of this scandal by the City of New York's Department of Investigation available at http://www.nyc.gov/html/doi/html/about/cases_bribery.shtml - and newspaper articles e.g. in the New York Times: http://www.nytimes.com/2008/06/07/nyregion/07crane.html?pagewanted=all&_r=0 and a recent article covering a broader scandal in construction-related inspections http://www.nytimes.com/aponline/2015/02/10/us/ap-us-bribery-investigation.html.

[317] See e.g. the following posts and articles: http://marginalrevolution.com/marginalrevolution/2014/08/ferguson-and-the-debtors-prison.html on the excessive use of fines as a budget funding mechanism – as well as the following: http://www.latimes.com/business/hiltzik/la-fi-mh-the-ferguson-crisis-20140821-column.html https://www.themarshallproject.org/2015/04/29/david-simon-on-baltimore-s-anguish http://www.governing.com/topics/public-justice-safety/gov-ferguson-missouri-court-fines-budget.html and on "civil asset forfeiture" and its abuse: http://www.vox.com/2014/10/14/6969335/civil-asset-forfeiture-what-is-how-work-equitable-sharing-police-seizure.

[318] See Ogus (2004), section *The Benefits of Corruption* for a discussion of some of the ways in which corruption enables inadequately (in particular: excessively) regulated economies to function nonetheless.

an economic perspective]" but "in practice, however, the creation of rents for the bureaucrats and politicians through regulation is often inefficient, in part because the regulators are disorganized, and in part because the policies they pursue to increase the rents from corruption are distortionary" (p. 3). Indeed, looking at a "cross-section of countries" they "do not find that stricter regulation of entry is associated with higher quality products, better pollution records or health outcomes, or keener competition. But stricter regulation of entry is associated with sharply higher levels of corruption, and a greater relative size of the unofficial economy" (p. 4). Overall, the research data shows that "better governments regulate entry less" and that "entry is regulated because doing so benefits the regulators" (p. 5). Djankov *et al.* further conclude that "the regulation of entry produces the double benefit of corruption revenues and reduced competition for the incumbent businesses already affiliated with the politicians" (p. 20) and that "entry is regulated more heavily by less democratic governments, and such regulation does not yield visible social benefits. The principal beneficiaries appear to be the politicians and bureaucrats themselves" (p. 27). In addition, in another paper (2006), Djankov *et al.* find that "results indicate that government regulation of business is an important determinant of growth and a promising area for future research. The relationship between more business-friendly regulations and higher growth rates is consistently significant in various specifications of standard growth models" (p. 4). Thus, abuse of business entry regulations appears to result in increased corruption, no visible social benefits, and reduced growth[319].

Even though business entry regulations are clearly not the focus of our research, these results can serve as a useful proxy for the relevance of investigations of how regulatory inspections are organized and conducted. Indeed, entry regulations, much like inspections, are primarily *procedures*, more than substantive regulations. What Djankov *et al.* have shown is that regulatory instruments, when excessively burdensome and indiscriminate, can have serious negative consequences on both the rule of law and economic growth. Similarly, abusive inspections can be expected to also lead to important negative results. In fact, investigating regulatory practices, and inspections and enforcement in particular, *in more details* can be expected to be particularly beneficial in terms of improving growth strategies. As Rodrik (2003) puts it (summarizing research by Kaufmann, Mastruzzi and Zavaleta), high-level reforms are often not enough, as shown in the case of Bolivia, where "the authors identify petty corruption, uncertain property rights, and inadequate courts as the source of problems". He emphasizes the need to "unpack "institutional quality" and show how aggregate indices or country averages can be misleading" (p. 14).

Preliminary conclusion

Concluding on the relevance of regulations, regulatory reform, and specifically of the improvement of regulatory instruments such as inspections to the complex issues of economic growth, social welfare and the rule of law is, to say the least, difficult. On the one hand, clearly, regulation and regulatory instruments are only one factor among many, and their short- and medium-term effects, at least, often pale in comparison with more immediate drivers (e.g. macro-economic policy). On the other hand, however, there is a converging body of research and evidence that points to their significance for long-term growth prospects, and to the harmful effects of "bad" (excessive, non-targeted, prone to arbitrary etc.) regulation. The limitations of data, as well as the complexity of the phenomena considered, means that absolute conclusions may be out of reach – but there seems to be enough ground to consider that making inspections work more effectively, efficiently and transparently is a worthwhile undertaking.

Supporting this view is one more angle that we have only alluded to so far, which is the relevance of inspection issues to trade. As we have shown above in the case of the US and the EU in particular, access to major markets

[319] On regulations having negative effects in terms of "barriers to entry", see also the pioneering work of Stigler (1971).

for important products such as food is increasingly subject to an exporting country's inspection systems being audited and found to be adequate (see also World Bank Group 2014 a). In other fields, countries have found themselves under pressure of potential boycotts, loss of trade preferences etc. because of glaring shortcomings in their occupational safety and labour law inspections (e.g. Bangladesh after the Rana Plaza disaster, or Jordan after labour abuses were revealed in the mid-2000s). Increasingly, a well-functioning inspection system is a pre-condition (or an important factor) for a country to avail itself of its trade opportunities. In all these areas, risk-based approaches are touted as an important way forward, making their study of real relevance to public policy

The point is not to come up with a "ready-made", "cookie cutter" approach, but to understand better the details of how inspections work, and with which results. As Rodrik (2003) writes, it is crucial to "go beyond simply asserting that "institutions matter" (…) [and] provide a richer account of where good institutions come from, the shape they take, and how they need to evolve to support long-term growth" (p. 12). Such work can support what has been called (Rodrik quoting Qian) "transitional institutions" (and, we would add, "transitional practices") that can be "more suited to the realities on the ground on both economic efficiency and political feasibility grounds" (*ibid*., p. 13).

The challenge, however, is to move from this recognition that "institutions matter" and that what matters are the details of how these institutions function, and with which effects. In this research, and particularly in the section covering practical cases, we will be attempting to look into the details of practices – but assessing the impact of these different practices is more difficult. As the brief selection of cases presented above shows, assessing the full economic impact of specific sets of regulations is a very difficult undertaking (assuming that it is even possible), and would require essentially an *ad hoc* study for each case, which would go far beyond the scope of this research. As a result, the only viable option for us was to select some *proxy indicators* for the economic impact of specific inspections and enforcement systems and practices.

As we have outlined above, the economic impact of regulation includes, crucially, *trust* (cf. Voermans, forthcoming, and the discussion e.g. of the history of food inspections). Unfortunately, quantifying the level of trust and its evolution would require specific surveys of market actors, that are not generally available. In order to look at the evolution of trust levels, a follow-up research would be required, looking for existing surveys and other data to try and construct indicators that can be compared over time and across jurisdictions. We were not able to attempt this within this research, but rather limited ourselves to anecdotal evidence suggesting higher or lower degrees of trust between jurisdictions, which we will consider in the overall conclusion. While this will obviously be inadequate to draw any strong conclusions, this may enable us to point towards directions for further research, and also have some preliminary indications of whether risk-based approaches appear adequate to provide the required level of trust.

Another side of the economic impact of regulation is (Djankov 2001, Kleiner 2006 etc.) more negative: barriers that limit market entry and reduce competition, procedures that give rise to corruption in various forms, costs that reduce profitability and productive investment etc. Many of these effects are, once again, difficult to measure – even though they may be the most significant. This is the case for instance of effects on competition, market entry, jobs etc. that were researched e.g. by Kleiner (2006). Direct administrative costs, by contrast, are relatively easy to capture. They are a key part of all Regulatory Impact Assessment (RIA) models that have been introduced since the late 1970s and have gained increasing acceptance since the late 1990s (cf. Blanc *et al.* 2015 pp. 48-49, OECD 1997, Radaelli 2007). One of the most widespread methods to measure direct administrative burden from specific regulatory procedures is what is called the "Standard Cost Model" (SCM), which is used in a number of countries (in the OECD and EU, but also developing countries), by

the EU itself, as well as by international organizations such as the OECD or the World Bank Group[320]. The SCM approach has been applied to inspections in various ways – either by relying on very detailed time measurements but a limited set of respondents (e.g. the "domain-focused" inspections burden measurement conducted in the Netherlands in 2007-2010)[321], or covering a larger (representative) sample of respondents but with far less detailed measurements (e.g. the calculations based on business surveys conducted by the World Bank Group e.g. in Kyrgyzstan, Tajikistan, Ukraine, Mongolia in the past decade)[322].

Unfortunately, administrative burden measurements are not an optimal measure. In many cases, they may count as a "burden" an inspection visit which, if conducted in a way that is transparent and focuses on compliance support, may in fact be experienced as a net positive by the business. In other cases, it may be that inspections appear to create relatively low burden in terms of what the SCM measures (primarily lost work time), but in fact create major barriers to business development through uncertainty, corruption etc. Administrative burden measurements, even though very frequently conducted, are clearly not measuring what matters most – and governments touting their success at decreasing burden sometimes miss the issues that most limit business developments. They are, however, far easier to conduct than other measurements, and are relatively frequently available.

In conducting this research, we have chosen to settle on an indicator that is simpler than aggregate administrative burden (as estimated through SCM calculations), but that in our experience is more reliable than an SCM based on a small sample[323], and can be an acceptable proxy for many other aspects of inspections: the overall number of inspections per business (combining both coverage – the percentage of businesses inspected in a given year – and frequency – the number of visits per inspected business). First, a high number of inspections is a very strong component of administrative burdens. It is very rare to have a country where very frequent inspections do *not* result in high burden – it would require extremely short inspection visits and, even in such cases, the aggregate burden remains significant[324]. Second, a high number of inspections is often indicative of an approach that relies primarily or exclusively on deterrence, and not on compliance promotion, and thus of inspection visits that are indeed perceived generally as a burden by businesses. Clearly, this indicator is not sufficient to indicate proof of negative economic impact, but there is sufficient evidence to suggest that it is generally an acceptable proxy for it[325]. We hope that, based on the preliminary findings of this research, there will be sufficient indications of the relevance of further research to support additional work, that would consider more closely the question of economic impact – being mindful of the fact that,

---

[320] See International working group on Administrative Burdens (2004) and Lundkvist (2010) as well as SCM Network (undated).

[321] These studies were not compiled in one general report, nor is there a general page presenting them. There were both baseline measurements and post-reform measurements. Some of the reports can be found at the following links: https://www.ilent.nl/Images/Eindrapportage%20Nulmeting%20toezichtlast%20vervoer%20over%20water_tcm334-318315.pdf https://www.nvwa.nl/onderwerpen/inspectieresultaten/bestand/26422/ https://www.ilent.nl/Images/0000%20Eindrapport%200-meting%20toezichtlasten%20domein%20overige%20chemie_tcm334-320054.pdf https://www.rijksoverheid.nl/binaries/rijksoverheid/documenten/rapporten/2010/04/22/eindrapportage-regeldruk-bedrijven-april-2010/voortgangsrapportage-april2010.pdf . All reports include the methodology.

[322] See section 4.1.b for a presentation of these surveys. Reports see e.g. World Bank Group 2009 (Mongolia), World Bank Group 2012 (Kyrgyzstan). Both reports include methodology and calculations.

[323] As inspections are not a universal procedure that every business has to go through, but rather can affect some businesses and not others. See e.g. Blanc (2009) on the limitations of SCM exercises for inspections when relying on a very small sample, and see 4.1.c for an illustration of how inspections can be concentrated on a limited sub-set of enterprises, meaning that having an unrepresentative sample could bias the results very strongly (example of Italy).

[324] This was the case e.g. in Kenya, where a 2010 survey conducted by the World Bank Group that around 90% of businesses were inspected each year, on average more than 5 times a year. The visits were mostly very short, so a strict SCM-type burden measurement would have given relatively low figures (though not *very* low) – but the burden was quite high in fact, because many visits were associated with corruption or harassment (report unpublished).

[325] See also LBRO (2010) for business perspectives showing both that inspection visits, when done in a supportive way, can be seen as more positive than burdensome, but also that very frequent visits are seen as a problem, and indicative of a poor business climate.

given the complexity of the phenomena observed and the interactions, full certainty of effects may not be achievable.

3.2. Promoting compliance: models, drivers, methods and issues

a. Models of compliance – deterrence, cooperation, responsive regulation – and beyond

i. A brief overview of compliance models

Introduction – the limits of compliance

Attempting to improve the effectiveness of regulation in achieving its intended effects, understanding the role of inspections and enforcement in this perspective, as well as the relative merits of different inspection approaches, all require to understand the links between rules and compliance, and the drivers that push people to comply. Of course, understanding and analysing compliance is not the same as explaining how outcomes in terms of public welfare are achieved, for there is not necessarily a direct link from compliance to public welfare (and there are in fact many situations where even perfect compliance will be insufficient to achieve the regulation's stated goals). We have discussed briefly above the question of the optimal precision of rules, which appears to be an impossible quest: only "target" technical norms would seem to give the "certainty" that what is required from the business corresponds to the intent of the regulation, and compliance *ipso facto* is equivalent to the desired result – but these norms leave business in complete uncertainty as to *how to* reach the desired result, and usually put inspectors in a difficult situation too, because of time-lags, third party effects etc. Given that *in practice* the vast majority of technical norms are "performance" or "specification", or a combination thereof, there is generally an imperfect match between compliance and intended regulatory outcomes. This mismatch is highest when regulation was inadequately drafted (because of haste or incomplete knowledge), and tends to increase as technological changes accelerate, or when third party effects increase (for whichever reason). This partial disconnect between compliance and outcomes (safety, public welfare etc.) is consequential for inspections, and one of the reasons some advocate for risk-based inspectors' discretion to be able to respond with more flexibility to the situation as it develops[326] – which, in turn, raises concerns from a rule-of-law perspective.

In fact, it has been demonstrated that even "target" rules are not immune to being "gamed", and emptied of their meaning – because it is impossible for rules to "target" everything that would be meaningful (not to mention the problems inherent in data collection). Bevan and Hood (2006) have thus shown how the "governance by targets and measured performance indicators" introduced by Tony Blair's "New Labour" government in Britain in the 2000s did not necessarily produce the expected results. Reported performance data appeared to show "notable improvements in reported performance by the English NHS" (National Health Service). In practice, however, there is substantial evidence that the system was gamed and that improvements were often "offset (…) [by] reductions in performance that was not captured by targets". While the authors rightly point out that none of the alternatives to this target system is "problems-free", their work

[326] See e.g. WRR 2013 (*op.cit.*)

clearly shows that we cannot expect that achieving targets automatically equates achieving the regulatory goals[327].

No form of rules appears to be immune to some form of gaming or evasion, e.g. through "creative compliance" that "uses formalism to avoid legal control" (McBarnet and Whelan, 1991). Indeed, "the combination of specific rules and an emphasis on legal form and literalism can be used artificially, in a manipulative way to circumvent or undermine the purpose of the regulation". We would add that, in our experience, such formalism that defeats the true intent of the regulation can just as easily be wielded by abusive officials (be they motivated by corrupt rent-seeking, or "simply" the enjoyment of arbitrary power) as by evading businesses. The proportion of creative compliance versus enforcement power abuse will depend on the characteristics of the country and of the regulatory interactions, but both are inherently possible in a system of rules, and quite difficult to fully avoid.

Going "beyond compliance" thus seems to be necessary in order to really achieve the full purpose of public policies, the full intended effects in terms of public welfare. Gunningham, Kagan and Thornton (2003) have devoted a large part of their research on the different degrees of "greening" of polluting industries to precisely this question of what could drive businesses to *exceed* regulatory requirements and engage on more significant and comprehensive pollution-abatement. At the same time, the authors point out why, in spite of its obvious limitations, so many still look to regulation (and enforcement) as the foundation for improvements (in the environmental field and elsewhere): "until the past decade or so, politicians, environmentalists, and scholars, observing the ongoing degradation of the environment in industrial societies, understandably assumed that the opportunities for such "win-win" investments were few and far between (…) and hence it has been assumed that legal coercion is necessary" (p. 21).

Indeed, even though rules cannot be designed "perfectly", and thus compliance with rules cannot fully ensure that regulatory goals are reached, rules and compliance still appear as a necessary foundation – necessary, though not sufficient. There remains a significant degree of causality between compliance and safety or other public welfare goals (as long as regulations are at least somewhat competently drafted and up-to-date), and in practice the shortcomings of the different types of rules will be somewhat alleviated by combining them ("specification", "performance", "targets" – as well as systems-based "to ensure genuine and long-term (…) improvements" – Tilindyte 2012 p. 17). This all matters to us because the primary purpose of regulatory inspections is, precisely, to *increase compliance*[328]. Of course, for inspections to be effective at this task, the drafting of the rules definitely matters[329] – but so do a number of other factors that we will now discuss.

Models of compliance - foundations

[327] Of course, the targets-based management of public health service provides that the authors study, while it is in some ways a form of "regulation", is quite distinct from the types of regulation we focus on in this research. What is relevant from our perspectives is that this shows that even targets-based rules cannot be automatically assumed to deliver the intended substantial outcomes.

[328] Or at least *one of the primary purposes*. There is also, in a different perspective, an "expressive interest of justice" (Hawkins 2002 p. 7), which demands as much as possible detection and punishment of rules violations – and detection requires inspections. The instrumental view of inspections and regulations corresponds to utilitarian values. Different sets of values (e.g. putting fairness and the rule of law first) will put a greater emphasis on the need to *enforce regulations* regardless of whether this is effective at achieving these particular regulations' purposes (cf. Morgan and Yeung 2007 p. 200, Ashworth 2000, Yeung 2004). We will return later to this question of values.

[329] Diver (1983) has proposed a set of "subcategories of potential costs and benefits" of different types of rules: *rate of compliance* (precise rules perform best), *over- and under-inclusiveness*, *costs of rulemaking* and *cost of applying a rule* (again precise rules tend to work better). The details of these criteria show how much depends on implementation, i.e. inspections. Thus, regardless of the type of rules chosen, and as long as they are more or less "fit for purpose", effectiveness will largely depend on the enforcement stage.

Many of the views held on public policy issues are based on assumptions, foundations, that are rarely even perceived, let alone questioned. Academics are not exempt from such "blind spots", with fundamental assumptions often remaining unchallenged for long periods of time. A common view underpinning (consciously or not) public demands for "more inspections", "more checks", "more enforcement" in relation to perceived risks (after an incident or in view of an "emerging risk") is that people comply with rules only if they are under supervision and there is a realistic threat of punishment for violations. This view is held even more widely and strongly in regard to businesses, which are seen by many as purely profit-driven – "amoral calculators". Business operators and owners are thus commonly held to be likely to comply if the costs of non-compliance are high, and punishment close to certain. The specific mistrust of businesses is often associated in the political field with radical left or anti-capitalist views, but the overall belief that people comply only under pressure and supervision is quite frequent in more conservative perspectives, so overall this view of people as reluctant to comply, and of regulation as requiring very strong enforcement to function, is held very widely and across the political spectrum (with different points of emphasis – but a shared foundation). Interestingly, such a perspective is also that of the compliance model proposed in 1968 by Gary Becker –which happens to be the first of the modern compliance models to have been formalized and still remains very influential[330].

This view, anchored in a pessimistic view of human nature (and understandably given credibility by the fact that crime and violations seem to be always recurring, and by human proclivity to estimate probabilities from negative experience and not from statistics – see e.g. Bennear in Balleisen, Bennaer, Krawiec & Wiener, in press), has been further reinforced by successive works attempting to model compliance based on neoclassical economics[331]. In these models, compliance is strictly based on maximisation of expected utility. The costs of compliance are weighed against the potential gains of non-compliance, minus the costs of possible sanctions multiplied by the probability of detection[332]. This model offers a convenient formalisation of the commonly held "pessimistic" view described above. The question is whether this model in fact describes observed behaviour accurately.

Defining this first compliance model is relatively easy – but there are several possible typologies of the other models. Scholz (1994) proposes a tripartition of what he calls "enforcement techniques and strategies" (that each correspond to a vision of what produces compliance): "deterrence strategy", "educational strategy" and "persuasive or cooperative strategy" (p. 425). He notes that the first is the "most familiar and best understood", "based on the assumption that regulated entities are amoral, and will not obey regulations unless given an incentive to do so". On the second, he writes that it "assumes that at least some noncompliance stems from the difficulty certain firms have with understanding regulations and implementing effective internal controls to prevent noncompliance" – an "educational strategy" does not "shy away from coercion" but rather uses it "to focus attention rather than to punish noncompliers", and it acknowledges potential negative side-effects of sanctions ("distracting" from some "recurrent problems", fostering resistance etc.). Finally, the third approach "assumes that firms perceive enforcement agencies as one of several important actors that the firm must deal with over the long haul" and that "firms develop principles to govern their relationships with all actors" (*ibid.*). Thus, while the first approach is squarely grounded in a deterrence model, the second corresponds to a model that introduces (and holds as fundamental) the issues of *understanding of rules* and of *capacity to comply.* The third approach is one that seeks to go beyond

[330] In our view, the influence of Becker's model (in spite of its limitations, which we will discuss below) can be traced both to its congruence with the commonly held "pessimistic" view of compliance, and its alignment with fundamental neo-classical economic assumptions (full rationality of market actors), and thus its being the most frequently taught compliance model in economics faculties.
[331] For instance, in the field of tax compliance, the works of Becker (1968), Allingham and Sandmo (1972) and Srinivasan (1973).
[332] This means that a maximal sanction of 1,000 EUR combined with a detection probability of 10% will result in an expected cost of 100 EUR – if the benefit from non-compliance is higher than 100 EUR, the person or business will choose not to comply.

compliance considerations – it should reduce the "inherent economic inefficiency of the regulations being enforced and the costs of monitoring and prosecution" (p. 441), and most importantly "maximize goal achievement rather than compliance" (p. 442). But the persuasive strategy also corresponds to a compliance model that sees firms as far more complex than the deterrence vision: "the persuasive strategy assumes that desired behavioral changes will only occur if they reflect the self-interest of the firm, just as the deterrence strategy does [but] the primary difference between the two (…) is that persuasive techniques appeal to a broader range of motivations (…), particularly the firm's concerns over its long-term relationships with the agency and other organizations such as unions, suppliers, purchasers, and the general public" (*ibid.*). Through his comments (pp. 425-448), Scholz suggests that all enforcement strategies have strengths and weaknesses, and that combining them has clear advantages but is far from easy because of contradictions between at least some of them (deterrence vs. education). Overall, he appears to be closest to a compliance model corresponding to the "persuasion" strategy – firms as rational actors with relatively complex calculations of costs and benefits, with deterrence being applicable to some situations and education to others, and a well-designed set of complementary strategies being optimal (thus suggesting that both "ignorance and lack of capacity" and "amoral calculations" drivers are relevant, either in the same firms or in different ones).

Another perspective is that put forward by Kagan (1994) of "legal enforcement style" – combining two perspectives, "the way officials assess compliance or noncompliance with regulatory standards" and "what officials do once they have decided that the regulated enterprise's actions constitute violations" (p. 387). Kagan then summarizes the different "styles" in a chart where the "enforcement style" can range from "inactive/unresponsive" to "active/responsive" and from "retreatists" to "legalistic" through "conciliatory" and "flexible" (p. 388). Kagan then attempts to connect these "enforcement styles" to "regulatory outcomes", which themselves can range from "excessively lenient" (ineffective) to "excessively stringent" (effective but with considerable negative side-effects, inefficiencies)[333], through an optimal "welfare-maximizing" range (pp. 388-389). These classifications are primarily based on a vision of intra-agency dynamics (attempting to understand why enforcement styles differ based on a combination of "legal design", "task environment", "political environment" and "agency culture" – pp. 390-391) – and on an economic perspective (looking for economically efficient outcomes, while acknowledging that "it usually is very difficult to determine whether agency enforcement decisions produce" them – p. 389). The underlying model for compliance is one which is based on rational calculations – a somewhat sophisticated vision of deterrence, including economic considerations for the viability of the approach (p. 398), and incorporating the "tit-for-tat", "responsive" approach to deterrence formulated first by Scholz (1984). The foundation remains one where enterprises are to be motivated for compliance through rational calculations, and the reason to not select a "strong" deterrence approach is only overall *public welfare* maximization, not the idea that that strong deterrence could be less-than-fully-effective at maximizing *compliance*.

In their influential work *Responsive Regulation* (1992), Ayres and Braithwaite[334] put forward a preferred model of enforcement as well as a vision of why businesses comply that is significantly different from the "deterrence model" and its variations. The authors formulate the fundamental debate as being "between those who think that corporations will comply with the law only when confronted with tough sanctions and those who believe that gentle persuasion works in securing business compliance" – with "most, although by no means all, regulators (…) in the compliance camp" and "most regulation scholars (…) in the deterrence camp" (p. 20). They add that many academics (of whichever ideological persuasion) "interpret this state of affairs as evidence of how captured the regulators are" (*ibid.*). Instead, they suggest that one can "strike some sort of

[333] The emphasis on deterrence has effects on inspectors' practice that lead to consequences in how they inspect, seeing the inspection more as a "case to be won" than as "problems to be solved", which tends to lead to poor cooperation, and can make actual detection and solving of real hazards less likely – see Bardach and Kagan 1982, pp. 80-81.

[334] Itself building on Scholz (1984), Braithwaite (1985), Braithwaite and Grabosky (1986) etc.

sophisticated balance between the two models", the question becoming "when" to use which approach (p. 21). The motivations for mixing or balancing the two approaches comes, however, from a somewhat different perspective than Scholz's (*cf. supra*) – indeed, while Ayres and Braithwaite also use the game theory perspective as a foundation for "tit-for-tat" enforcement (p. 21 and pp. 60-81), they combine it with a greater attention paid to the "mixed motives" for compliance[335]. Based on empirical work, they propose "alternative motivational accounts" to the vision of "the firm (…) [as] a unitary actor concerned only with maximizing profit" (p. 22). First, corporate executives value "a good reputation" and care "deeply about the adverse publicity", viewing "their personal reputation in the community and their corporate reputation as priceless assets" (*ibid.*)[336]. Second, "corporate actors are not just value maximizers – of profits or of reputation", but also act according to values (ethics, social responsibility etc.). In practice, "there is evidence of economically irrational compliance with the law" (pp. 22-23). The authors are realistic about the strength of values-based motives, and recognize that they will often not be stronger than profit-based motives, but insist on the need to recognized that, in a significant proportion of cases, they are (pp. 23-24). Third, another key aspect that Ayres and Braithwaite emphasize is that "firms are not monolithic" and that "not all of the relevant actors have the same interest in profit maximization as those at the top may have" – there can be, in any organization, "law-abiding constituencies" (p. 33). Overall, they propose a model of compliance that emphasizes *complexity* and *multiplicity*: several drivers, several groups, several motives inside a same person or company. Depending on the business being considered, and on the situation, various combinations may arise, and the resultant profile may be more or less profit-maximizing, more or less ethical – and Ayres and Braithwaite see it as "responsive" enforcement's purpose to reinforce the compliance-maximizing forces and weaken the others, meaning that theirs is also a view of *dynamic* compliance drivers. Crucially, this includes the possibility that some enforcement actions that would make sense from a deterrence perspective would be counter-productive by decreasing intrinsic (values-based) compliance forces[337]

The emphasis on the complexity of factors leading to compliance, and the view of corporations as multiple rather than unitary, are further developed in the work of Gunningham, Kagan and Thornton (2003)[338]. They emphasize the importance of "endogenous" factors within corporations: "more generally, "new institutionalism" theories of organizational behavior reflect findings, as summarized by Mark Suchman and Lauren Edelman, that: 'institutional factors often lead organizations to conform to societal norms even when formal enforcement mechanisms are highly flawed. Frequently cited institutional influences include historical legacies, cultural mores, cognitive scripts, and structural linkages to the professions and to the state'" (pp. 22-23)[339]. More fundamentally, they propose a new, broader concept, instead of a narrow and unilateral causality

---

[335] Which, in turn, is also reflected in Scholz's later work (1994), where he includes the complexity of motivations in grounds for the "persuasive strategy".

[336] Though Ayres and Braithwaite do not directly qualify this claim, they also do not suggest that it is universally true. We would add that it clearly is not, the financial crisis that started in 2008 having revealed the depth of inadequate corporate behaviours in the financial sector, and the apparent weakness of the "adverse publicity" driver for many of its executives (though on the other hand the many *op eds* written by banking executives, e.g. in the US, seems to show that adverse publicity is still something that they strongly resent – it may just not be enough to overcome other drivers).

[337] A point also noted by Scholz (1994) when noting that deterrence strategies can contradict education ones.

[338] Itself building on earlier work such as Gunningham and Grabosky 1998, Gunnigham and Johnstone 1999.

[339] A note here is needed: in this book, the authors specifically look not only at what makes firms comply with regulations, but also what makes them go *beyond*, i.e. improve environmental performance above and beyond regulations. That said, many of the factors at play are essentially similar when it comes to complying with or when it comes to *exceeding* mandatory norms. They build a typology of reasons for firms to go *beyond* compliance (p. 24) – it includes "win-win measures" ("sometimes the firm [invests in nonrequired methods] because [they] are more cost-efficient than those required by the rules, and sometimes because they feel it is "good business" to develop co-operative and mutually trusting relationships with regulatory officials" - p. 21), "margin of safety measures" (ensuring that compliance is always guaranteed even when there are variations in production conditions), "anticipatory compliance measures" (avoiding costly upgrades/retrofits when regulations change by building equipment "one or two steps ahead") and "good citizenship measures" (which improve the image of the company e.g. with consumers). To some extent, all these different cases correspond to a broad vision of "profit maximization", based on a far larger consideration that the narrow cost of compliance vs.

of "drivers": "in the course of our field research, we came to regard the concept of "drivers" as somewhat impoverished. It implies the existence of independent, unidirectional, and unambiguous pressures, whether from regulation, communities, or markets, which impact upon corporations with sufficient force that they react to them. Yet we found that these external factors, rather than being independent, often gain their force through mutual interaction; that far from being unambiguous, the responses they demand are often unclear; and hence that they do not operate unidirectionally" (p. 35). By contrast to the over-simplifying notion of "drivers", "the concept of a license to operate (…) captures the complexity of the relationship between the regulated enterprise and key stakeholders in a way that the concept of "drivers" does not" – it "encapsulates the extent to which various stakeholders can bestow or withdraw privileges from a company" and "that business is dependent upon" these stakeholders' relations – it also means that the relationship "is an interactive one" – it includes "the regulatory license, the economic license, and the social license" (p. 36)[340]. Among the different forces at play, the authors find a combination of different dimensions, which correspond to their vision of three different "licenses": "the external pressures that push enterprises toward improved environmental performance can be divided into three broad categories: economic, legal and social" (p. 35). This tri-dimensional vision already represents a considerably higher level of complexity than the narrowly economic one, and a model that appears better suited to a diverse reality.

Other authors have researched and underlined the importance of a fourth category of "driver" or "pressures": psychological factors. As Hodges (2015) puts it: "the science of cognitive and behavioural psychology has undergone revolutionary development in the past few decades" – but, he adds, "the findings have not been noticed by many legal theorists" (p. 2). In fact, both among legal scholars and economists, there is a significant group of authors who have built research and models on the basis of these advances in psychology – but it remains true indeed that the majority of scholars tend to ignore them and to rely on far cruder models and visions of human motivations. By contrast, "psychology posits decision-making that is based on multiple factors other than costs and benefits" (regardless of whether one speaks of monetary benefits, or "immaterial" ones such as reputation etc.) (*ibid.*). One of the most important compliance models based on psychological insights is commonly called *procedural justice*, and has been developed and exposed in particular by Tyler (1988, 1990, 2003 *et al.*)[341]. Importantly, this approach does not negate other insights on economic or social drivers of legal compliance, but rather subsumes them in a more comprehensive vision – while suggesting that "psychological" and "social" factors (ethics, legitimacy, procedural justice) may be stronger than "economic" (deterrence) ones[342].

The compliance model developed by Tyler views legal compliance as driven by a combination of motives: rational calculation (deterrence) being one, along with moral values, social norms, legitimacy and procedural

---

probability of detection and potential sanction calculation that is at the core of the "deterrence" model, but still mostly predicated on the same "amoral" logic.

[340] The authors use of the word "license" also corresponds to their studying a population of businesses that is, in fact, subject to precisely this form of regulatory instrument (prior approval). Their study population is made of generally large companies, large facilities, with a very high environmental impact – and correspondingly strong regulatory attention. Even though this means that not all of their findings or lessons are applicable to other fields, their view of the complexity of compliance (and "beyond compliance") factors appears relevant far beyond this study population and aligns well with the findings of other research.

[341] The notion of "procedural justice" can be found already in Rawls's *Theory of Justice* (1971). The development of this notion in a legal compliance perspective is due e.g. to the work of Leventhal (*Fairness in Social Relationships*, 1976) as well as Thibaut and Walker (*Procedural Justice*, 1975). Tom Tyler has been one of the researchers leading the development of this notion specifically applied to the question of "why people obey the law". E. Allan Lind was another early proponent of this vision, which is now supported and used by a growing number of scholars – and practitioners.

[342] The distinction between "psychological" and "social" is not necessarily an easy one – just as biology and physics, the two attempt to describe and explain the same reality, but at different levels of detail or "granularity". We will qualify as "psychological" the factors that are related primarily to the internal views and thinking mechanisms of the individual, as "social" those that primarily involve group values and behaviours, and as "economic" those that (while of course anchored in psychological mechanisms and social values too) correspond to (neo-)classical economics' emphasis on "amoral calculations" and pure rationality.

justice (which, in turn, reinforces legitimacy). Procedural justice is a term that corresponds to authorities treating those subject to them in a fair manner, irrespective of the outcomes of the decision-making process[343] - and this "fair process" is defined as one that combines consistency (of treatment, criteria etc.), impartiality (or at least perceived best efforts to be impartial), ethical behaviour (including civility of persons in a position of authority) and adequate representation (i.e. giving a "voice" to the person affected by the procedure)[344]. Within this model, legitimacy (of authorities and rules) is seen as the real foundation of compliance (see Tyler 1988 pp. 19-70), and procedural justice as the instrument through which such legitimacy and compliance can most effectively be increased. The central importance of procedural justice in this model lies therein that it appears to be the factor that can be most easily strengthened, as well as individually possibly the most potent one. Indeed, deterrence appears very costly, and relatively weak. Moral values are built from childhood, and difficult to alter. Social norms (prevalent behaviour in a given group) are a complex product of various influences, and thus usually can only be altered gradually. By contrast, legitimacy of public authorities (and of the rules they impose) appears strongly influenced by procedural justice[345] – which also has its own positive compliance effect. Thus, procedural justice appears overall as the most important factor in increasing compliance. It is also *relatively* easier for authorities to influence since it depends directly on how they behave in relation with those affected by their actions (citizens, businesses etc.). These findings rely on several decades of research on criminal matters, and on interactions between citizens and authorities (police and courts in particular), as well as some more recent (and so far less extensive) research on regulatory dealings (Lind and Maguire 2003) and public services interactions (van den Bos, van der Velden, and Lind 2014). At its core, the model states that deterrence does play a role in fostering compliance (i.e. deterring crime), but that it tends to have an effect that is quite limited, except if considerable resources are expended so at to make the probability of detection really high. On the other hand, process-based factors appear to play a crucial role in determining sustained attitudes in respect with laws and regulations, and with public authorities.

Tom Tyler summarizes deterrence's impact and limitations as follows (2003): "studies of deterrence (…) point to factors that limit the likely effectiveness of deterrence models. Perhaps the key factor limiting the value of deterrence strategies is the consistent finding that deterrence effects, when found, are small in magnitude. (…) A further possible limitation of deterrence strategies is that, while deterrence effects can potentially be influenced by estimates either of the certainty of punishment or its severity, studies suggest that both factors are not equally effective. Unfortunately from a policy perspective, certainty more strongly influences people's behavior than severity, and certainty is the more difficult to change. (…)To influence people's behavior, risk estimates need to be high enough to exceed some threshold of psychological meaningfulness" (p. 302). This means that, in practice, deterrence is impossible to achieve in most cases: the resources required would be far too high (in a world of limited resources, society cannot commit enough resources to deterring violations in each and every regulatory field), and the intrusion on privacy and limitations of individual freedoms would be far too high. Tyler cites murder as a key example: on this topic, society has allocated enough resources that indeed there is a real deterrence effect – but achieving similar intensity of enforcement in all other fields is

[343] The fairness of outcomes corresponds to distributive justice – which is often difficult to assess independently or objectively, meaning that the perception of distributive justice tends to vary from one person to the next and (when there is a conflict) may tend to be zero-sum: what one perceives as a fair outcome is seen as unfair by the other. By contrast, procedural justice can be perceived by *both opposing parties* as high, since it relates to characteristics of the process, and not to the outcome.

[344] See e.g. Tyler 1988 pp. 136-139.

[345] We are simplifying here (on purpose and to make its main points clearer to the reader) the complex model developed and extensively tested by Tyler (1988). In this model, Tyler tests a number of cross-relations between different factors or drivers, and there is evidence of multiple influences on legitimacy, including not only procedural justice but also (perceived) distributive justice. This influence, however, is consistantly found to be *weaker* than that of procedural justice (a finding strongly confirmed by van den Bos, van der Velden, and Lind 2014) – and, in addition, consistantly increasing perceived distributive justice is very difficult, given the conflicting views of it that co-exist (see previous note). Thus, procedural justice appears not only as the strongest, but also the most realistically "improvable" driver of legitimacy. See Tyler 1988 (pp. 106-109 in particular) as well as Bottoms and Tankebe 2013.

impossible. In addition, deterrence approaches "are not self-sustaining and require the maintenance of institutions and authorities that can keep the probability of detection for wrongdoing at a sufficiently high level to motivate the public" (p. 304).

By contrast, process-based approaches aim at increasing the legitimacy of rules and authorities by improving the level of fairness as perceived by citizens. The focus is not primarily on "distributive justice" (i.e. having *outcomes* that are deemed fair) – although this also has been found to have a significant impact on compliance, it is significantly less strong than the process effect, and in addition it is in practice impossible to reach decisions that would satisfy everyone. Rather, the emphasis is on "procedural justice". In the words of Tyler (2003), who has been one of the key proponents of this approach for several decades: "The procedural justice model involves two stages. [First,] public behavior is rooted in evaluations of the legitimacy of the police and courts. (…) In other words, people cooperate with the police and courts in their everyday live when they view those authorities as legitimate and entitled to be obeyed. [Second,] the antecedents of legitimacy. The procedural justice argument is that process-based assessments are the key antecedent of legitimacy (…). In this analysis, four indicators – summary judgments of procedural justice, inferences of motive-based trust, judgments about the fairness of decision making, and judgments about the fairness of interpersonal treatment-are treated as indices of an overall assessment of procedural justice in the exercise of authority" (p. 306). Crucially, a considerable body of research has shown that the effect of procedural justice appears *significantly stronger* than that of deterrence[346].

The procedural justice effects are found in many fields and settings (mediation decisions Lind et al. 1993, dismissal from employment Lind et al. 2000 etc.). What also matters is that procedural justice, and the legitimacy it fosters, are long-term drivers of compliance, and largely self-sustaining (at least they do not require an *increase* in resources – but a change in behaviours and approaches)[347]. The changes involved in how authority is exercised are, however, significant compared to what is the practice in many cases. Quoting Tyler (2003) again, the key conditions needed to achieve a procedural justice effect are: "that decision making is viewed as being neutral, consistent, rule-based, and without bias; that people are treated with dignity and respect and their rights are acknowledged; and that they have an opportunity to participate in the situation by explaining their perspective and indicating their views about how problems should be resolved" (p. 300-301).

The validity (or lack thereof) of different compliance models is in no way a purely "academic" question – since it provides the foundation for different inspections and enforcement approaches. A "classic" deterrence-based approach (where increasing probability of detection or severity of sanctions are seen as equivalent) will lead to the use of punitive sanctions or damages (in tort cases), whereas a deterrence-based view that takes into account research suggesting that people react more to probability than to severity will try and increase inspections coverage and at the same time refine targeting (e.g. by doing at least some basic "risk-based" targeting, looking for higher probabilities of violations if not magnitude of potential effects). By contrast, an approach that takes a more complex, multi-factor view of enforcement will be quite different. It will consider alternative approaches to promoting compliance (in particular education, guidance, opinion-forming), it will pay attention to the importance of ethical behaviour of inspectors and "procedural justice" more broadly. It will also look at the potential adverse effects of excessively frequent, burdensome inspections, or of enforcement seen as disproportionately severe. Indeed, if their negative procedural justice effects were to be higher than their deterrence effect (something which is a distinct possibility in such a model), then the net compliance effect of more inspections and stricter enforcement may well be negative. We will come back

[346] See Tyler 1988, 2003 – Hodges 2015 *et al.*

[347] On this point, see e.g. Tyler 1988 p. 107 (procedural justice acting as a "cushion of support when authorities are delivering unfavourable outcomes", as well as Tyler 2003 p. 283 etc.

further in this research on the evidence concerning different compliance models, but we can already say that, evidence notwithstanding, these different models have very concrete "real-life" effects, as different inspectorates across the world base their operations on very different visions.

ii. Mapping the foundations of compliance – economic, psychological, social, cultural

In summary, theoretical accounts of compliance and research-tested models have gradually moved away from a narrow, deterrence-based vision to a more complex, multi-factor model – or one could also say that such a complex vision has long existed, but has gradually gained ground against a once-dominant deterrence model. Indeed, the deterrence model appears overly simplistic – applied to business regulation, "it assumes that all businesses make all decisions based solely on objective economic rationality, weighing all costs and benefits in financial terms. It is further assumed that an organisation can be treated as a single entity, and that it can control the behaviour of every person and decision that is taken" (Hodges 2015, p. 3). Rather, firms are made up of many individuals, and human decisions and behaviour are shaped by their "cognitive development and "moral understanding", their "sense of justice", as well as "exemplars of a social norm or custom" (*ibid.*, pp. 15-16). Crucially, decisions are more often taken on the basis of the "fast heuristic approach", which "involves impulsiveness and intuition", than using the "slower system that is capable of reasoning [and] is cautious" (*ibid.*)[348]. Thaler and Sunstein have shown the importance of heuristic biases (1998, pp. 19-39) in our decisions. For all these reasons, effectively promoting compliance appears to require an approach that combines a number of drivers or dimensions.

We have seen that the number and categorization of such drivers varies between authors. Hodges (2015) sees "three primary motivations for explanations of law-abidingness in humans": "fear of detection and punishment", "fear of humiliation or disgrace" and "internalized sense of duty" – the latter being in turn influenced by "internalised moral values", "processes by which the rules are made and applied" and the alignment (or lack thereof) of "the rules and culture of the group(s) to which the individual belongs (…) with the norms that are sought to be applied by society" (pp. 16-17)[349]. In addition to these, we would also underline the importance of *capacity to comply*: both the knowledge and information aspect (emphasized e.g. by Scholz 1994, *cf. supra*) and the material side of compliance (technical capacity and feasibility, and cost of compliance)[350].

---

[348] Cf. Tversky, A. and Kahneman, D. (1974), "Judgment under Uncertainty: Heuristics and Biases" 185 no 4157 *Science* 1124-1131 – as well as Bennear in Balleisen *et al.* (in press), and Sunstein and Thaler (2008) pp. 19-39.

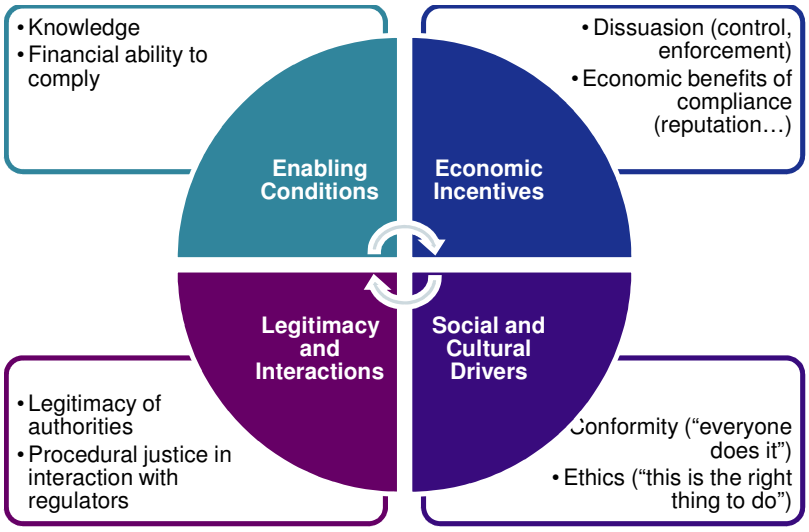
[349] See also Kagan, Gunningham and Thornton 2011 (p. 37): "Sociolegal explanations of law-abidingness among regulated business enterprises, as well as among individuals, point to three basic motivational factors: fear of detection and legal punishment; concern about the consequences of acquiring a bad reputation; and a sense of duty, that is, the desire to conform to internalized norms or beliefs about right and wrong"

[350] See e.g. Winter and May 2001. This aspect is also covered under the "economic" side of the "license to operate" concept outlined by Gunningham, Kaghan and Thornton (2003) – see also Ogus (2004) on how corruption can help an economy to function in spite of excessive, inefficient regulations. Note that Ogus takes a limited example (procedural regulations). In our experience, the effect is even stronger when substantive regulations are "impossible to comply with" given prevailing technical and financial conditions: rather than most businesses just shutting down, either implementation of the rule has to be scaled back by the regulatory authorities, or corruption will enable businesses to function nonetheless, but at a significant cost (and profit for corrupt officials) – in both cases, compliance will be non-existent (or close to it). For an example of non-implementation of an unrealistic rule, see the example of the constantly pushed-back full implementation of accessibility rules for handicapped people in France, Eliakim (2013) (chapter "Maintenant, ils regrettent…). For a similar example on lifts regulation in France, see Blanc *et al.* (2015) p. 8 and Eliakim (2013).

Drawing on these different streams of research and scholarship, and on inspections and enforcement practice, we would propose to categorize the different *foundations*[351] of compliance in four groups[352]:

- Enabling conditions: knowledge and understanding of rules, financial and technical ability to comply without putting the business viability in jeopardy
- Economic incentives: deterrence (probability of detection primarily, amount of potential sanctions as a secondary aspect – and also risk of reputation loss), potential economic benefits of compliance (increased reputation leading to improved market position, or compliance investments resulting in higher productivity, reduced losses or any other economic benefit)[353]
- Social and cultural drivers: group conformity (other members of the group and/or "models" behave in a compliant way), group ethical values (values of the cultural group the person belongs to are aligned with the values of the regulation and/or values of the cultural group posit legal compliance as an absolute good)
- Legitimacy and interactions – individual psychological drivers: legitimacy of authorities (influenced by social and cultural drivers, but also directly by personal, individual experience), procedural justice (or lack thereof) experienced in interactions with authorities, regulators.

Evidently, these categories are only useful as a device to help clarify and make sense of the complexity of compliance. "Enabling conditions" have economic aspects, "legitimacy" and "values" are both social and psychological, etc. Rather than four separate categories functioning in isolation, it is useful to see these as part of a circle of "contextual elements", all of them interacting with one another. Visually, one can represent it as in the following scheme.



An illustration of the "compliance foundations circle"

[351] A term that is broader than "drivers" in that it also includes elements that are rather "pre-requisites", "enabling" factors.

[352] Of course, many different categorizations are possible. Parker and Lehmann Nielsen (2011) for instance propose: "four main conceptual themes or sets of independent variables of interest in explaining compliance: motives, organizational capacities and characteristics, regulation and enforcement, and social and economic environments (or institutions)" (p. 5). We consider that having "regulation and enforcement" as one of the variables is too broad, for instance, and consider "institutions" in their different aspects under several headings. Each typology will have different strengths and weaknesses, and will depend on the focus of the author(s).

[353] The *complexity* (often under-estimated) of this group of "economic or material motivations that influence businesses to comply (or not) with regulatory dictates" is underlined by Simpson and Rorie (2011, p. 59): "our discussion acknowledges the importance of micro and macro distinctions and the linkages between organizational members and the company as a whole".

It is worth emphasizing that this is not a scientifically-grounded typology, where the different categories would be very tightly defined. Rather, their relative porosity reflects the inherent complexity of the behavioural processes that result in compliance, and the many interactions between all the factors at play. The intent is to have a typology that is above all practice-oriented: if indeed the factors described are found to be significant (and we will discuss below the question of their respective importance), then an effective and efficient inspection and enforcement approach has to try and address all of them comprehensively, paying attention to potential side effects, trade-offs, and attempting to find the "optimal" mix of tools[354].

In sharp contrast with some theories' pretentions to definitely "explain" human interactions, we think it behoves scholars to be modest and accept the limits of humans' ability to understand themselves[355]. Whether these limitations are inherent or temporary, we clearly are nowhere near reaching the same success in understanding *and transforming* behaviours as we have had in regard to natural, physical phenomena. Accepting and understanding to some extent the complexity of forces at play and of interactions may be a first step. In this perspective, breaking down artificial barriers between different theories is a first, useful step. In agreement with Hodges (2015), we see the combination of procedural justice studies, behavioural economics, as well as the more sophisticated analyses of "deterrence" effects as different aspects and angles of the same attempt to make sense of human behaviours. Whether one looks at "compliance", at "beyond compliance", or targets "behaviours" more broadly – the drivers, conditions, foundations are all essentially the same.

b. Challenges in understanding compliance, and promoting it

i. Determining the relative strength of compliance drivers: a difficult quest

Considering that there are several competing compliance models, and different perspectives on the relative importance of compliance drivers or factors, and that these different models and drivers suggest sharply diverging enforcement approaches, attempting to assess the relative strength of these different drivers is very important. It is also very challenging – at least if one wishes to have definitive certainty, or close to it. Many studies have attempted to test the effect of different approaches, in particular deterrence, but also (for a smaller number) procedural justice, education etc. Tyler (1988 *et al.*), in particular, has attempted to disaggregate the effects of different drivers, while testing a procedural justice model of compliance. Still, we would argue that none of the studies is fully conclusive, and that it is hence not surprising that many yield partly or fully conflicting results. All of these studies have their limitations: they cover typically one (often small) jurisdiction, and one legal or administrative field. There may be a number of reasons why the effects found differ between locations, topics, groups affected etc. There are also considerable issues with data

[354] As Parker and Lehmann Nielsen (2011) put it, "understanding and explaining 'compliance' (…) requires mapping, understanding and testing the interactions of a complex range of factors and processes" (p. 8). There are, of course, many different ways to consider and categorize existing research on and scholarly accounts of compliance. Parker and Lehmann Nielsen see the field as divided between "objectivist research aimed at building and testing theories" that look at "what 'procudes' compliance" – and "interpretive understanding of organizational responses to regulation, and of the processes by which compliance is socially constructed" (p. 3). If we had to choose, we would locate our work in the first group – but the authors quickly add a note that there is "creative dialogue" between the two, and that many scholars "use both styles" (p. 4).

[355] Tyler's *Why People Obey the Law* (1988), precisely because it attempts to capture all the different (and possibly conflicting) drivers, is a good example of such modesty and inclusiveness.

quality, reliability, representativeness in many studies – and with the meaningfulness of respondents' responses to "qualitative" questions.

We are not the first to point the limitations and difficulties of data on compliance. For instance, Kagan, Gunningham and Thornton (2011) write that "the regulatory agency databases that researchers use to measure noncompliance vary in quality, while researchers who rely on those databases often differ in what they treat as significant noncompliance" – but we think this is an understatement of the problem. Lehmann Nielsen and Parker (2011) go somewhat further in stating that: "to the extent that data are available from individuals inside firms or from records collected by regulatory agencies, the data will be filtered and biased according to what those who collected it saw as relevant and important to compliance and what they see as socially and politically desirable to share with the researcher" (p.6). Similarly, the conclusions drawn by May and Winter (1999, 2011) on different "enforcement styles", and their respective effects on compliance, while very interesting, are subject to caution given the limitations of the data they use (which they partly acknowledge). Their 1999 study relies on surveys asking respondents to rank enforcement styles on a set of criteria (which already can incorporate a significant amount of respondent bias, as with every "qualitative" questions) – and then combines this with questions where inspectors are asked to assess the effectiveness of their own actions. As the authors write, "we relied on municipal inspectors' reports of the effectiveness of their enforcement efforts" (2011, p. 234). While the authors do grant that there may be concerns with the objectivity of such a data source, they still consider it as fundamentally valid[356], and draw important conclusions from their research, in particular that "the effects of formalism [in enforcement style] were positive and somewhat stronger when awareness of rules was low. In such circumstances the use of formalism gives regulatees more certainty about what is expected from them" (p. 235). While the conclusion may well be valid, it remains a distinct possibility that more "formal" inspectors may also, for a variety of reasons, *consider* their own actions to be more successful – and thus, different perspectives may build an inherent bias in the data.

We would argue that self-reported levels of compliance, whether reported by businesses themselves or by inspectors, are highly problematic. The former may have an imperfect understanding of what full compliance would be, and a reluctance to report fraud and violations. The latter have a number of incentives to report compliance levels that may differ from reality (not necessarily better – policy priorities may also mean that reporting worse compliance than actual makes career sense), and also of course never have a full view of the level of compliance in any given business, even one that they inspected – and by definition have no information on non-visited businesses. While one may assume that the "imperfect information" issue may be relatively constant, and thus not skew evaluations of *relative* effectiveness[357] (or skew them only in a limited way), the same is not true of pressure from inspectors' management, policy makers etc. May and Winter (2011) in fact acknowledge the importance of superiors and politicians in how inspectors' enforcement style varies, for instance, even though they also find this effect to be variable and often limited (pp. 230-232) – there is no

[356] See May and Winter (1999): "Our measure of the effectiveness of enforcement actions in bringing about compliance is based on assessments made by the main municipal inspectors. Each was asked to rate on a 10-point scale the "total effect of the municipal supervision of farmers' pollution of water resources in relation to making farmers comply with regulations governing livestock." The end points for the scale ranged from "no effect" to "has caused all farmers to comply." Municipal inspectors generally report high degrees of effectiveness of their actions (…) Inspectors have some incentives to provide rosy estimates of their effectiveness; if nothing else, to look good. However, the Danish Environmental Protection Agency regularly requires the municipal inspectors to make reports about farm inspection for which inconsistent reports over time are evident. (These reports are one source of our data.) Given these considerations, we presume that the inspectors' reports of enforcement effectiveness provide reasonable measures of relative differences among municipalities. However, we recognize that they may not provide accurate assessments of absolute levels of enforcement effectiveness or of compliance" (pp. 635-636). The authors cite other studies that report inspectors' own assessment of compliance to be accurate. Clearly, it is *possible* that these are indeed accurate – but it is far from certain. Inspectors may well have understood how to "game the system" and consistently report "better than actual" outcomes (cf. Bevan and Hood 2006).

[357] Although more qualified and professional inspectors tend to be better at detecting violations, as May and Winter (2011) also note.

reason to assume that this influence does not also extend to reporting of compliance levels. In fact, extensive research in the field of crime and law enforcement has repeatedly shown major issues with the way police forces register and report crime levels (including under-registration of crimes that the police would be unlikely to be able to solve, so as to increase the rate of success – or systematic enforcement against petty crime in order to make "activity" statistics look up, thus making it appear as if there were a surge in some violations, etc.)[358]. There is no reason to believe that such problems are not also present in regulatory enforcement. Indeed, Bardach and Kagan (1982) have shown that, when inspectorate management emphasizes a "looking tough" approach and penalizes inspectors who appear to have lower activity and enforcement numbers, this mechanically produces a more legalistic, more "aggressive" enforcement practice (pp. 76-77), with considerable side-effects (what the authors call "unreasonableness"), without this reflecting on the real, underlying level of compliance and safety.

For all these reasons, we believe it makes sense, in order to assess the effectiveness of inspections and enforcement approaches and styles, to look at actual *outcomes* and not at whichever compliance levels are reported. Obviously, there are major issues also with this approach (in particular, the difficulties in attributing variations in extremely complex phenomena to different causes), and we will come back to this in the third section. In the meantime, and in spite of the limitations outlined above, considering the evidence from existing research is a crucial step in order to provide a sound foundation for inspection practices. We will attempt to do this briefly, focusing on the most significant results, and assessing whether some trends can be more or less reliably identified.

Two fields of law and regulation have been the object of most studies of compliance and its possible drivers: tax regulations, and interactions with police and courts ("law and order" issues broadly speaking, and not only criminal justice). While there is no comparable set of quantitative studies on other areas (environmental or occupational regulations compliance, for instance), there is good reason to assume that findings from these two spheres can extend to other fields too. Indeed, in neoclassical compliance models, the cost-benefit calculations are assumed to extend to any kind of regulation as well. From our perspective, taxes and "law and order" issues have the benefit of covering very different types of regulations – complex for tax and simpler for "law and order", applying only to individuals for the latter and also to businesses for the former, etc. That they have been most studied is a function both of their very strong relevance to society (very fundamental fields of state regulation), and of the relative ease with which compliance and non-compliance can be measured (quantitatively in tax, and with simple questions in terms of law and order – whereas environmental or occupational regulations, for instance, would entail many different questions and compliance could be partial, with difficulties in rating it). While neither of these fields is the core focus of this research, there is reason to think that findings in these spheres can be transposed to others[359].

Evidence from tax compliance studies

If we thus accept that we can generalize the findings from tax compliance and "law and order" studies to other fields of regulation, there is a significant amount of evidence *against* the view of people and businesses as complying only on the basis of fear and rational calculations. To quote from an important study reviewing and

[358] There is a vast amount of literature on this issue – the reader can refer e.g. to Skogan 1975, Smith 2006.

[359] Tyler (2011) himself considers the application of his findings to business regulation writing (p. 78): "deterrence mechanisms of the type being widely used are usually less effective than is generally believed, and are particularly unlikely to be optimal approaches to regulating the actions of those who work in business settings. In contrast, research findings suggest that efforts to build a value based climate of rule following are a promising approach that is likely to lead to more widespread voluntary acceptance of, and deference to, workplace rules and policies. (…)Studies find that the primary factor shaping legitimacy, morality and rule adherence is the procedural justice that employees experience in their workplace". We will come back to this question of the "ethical" workplace, which is also the focus of Hodges (2015).

summarizing several decades of research on tax compliance (Kirchler 2007) "empirical research consistently shows that the rational model is not working as neoclassical economics had intended[360]". Kirchler, in this study, goes through all the conflicting evidence put forward by a number of studies in different countries, some in a laboratory setting, some based on surveys, some others looking at actual tax data. Most show a stronger effect from audit frequency, a few from higher fines (though from a model perspective they ought to be equivalent), some show no effect or an adverse effect (more audits and/or higher fines leading to *decreased* compliance) – and in all cases the effects are small. Among the most interesting findings from our perspective are that "oppressive tax enforcement and harassment of taxpayers seem to increase tax resistance, as does discontent with the delivery of public service[361]" – and that another study[362] "yielded neither a significant audit probability effect nor significant effects of fine and tax rates, whereas trust in the legal system and direct democratic rights proved to be highly significant determinants of tax morale. These findings prove that perceived procedural justice as described above is a crucial determinant of citizens' voluntary cooperation, whereas in a system perceived as treating citizens unfairly, cooperation must be enforced by coercion".

Overall, Kirchler summarises the key findings as follows: "there are many explanations of why probability of audits and fines does not have the predicted high effect on tax compliance. First of all, the assumption that taxpayers are trying to avoid taxes if it is in their benefit must be doubted. Various studies in different countries use different methodological approaches to show that a vast majority of citizens are willing to pay taxes and do not seem to undertake economic decisions under uncertainty in order to maximise income. Most taxpayers seem to take for granted the legitimacy of the tax system and its overarching objectives". Even to the extent that audit probability and fear of punishment do play a role, their effects are mediated by the values of the taxpayers: "individuals generally make poor predictions of the probability of audit and magnitude of fines from tax evasion. Moreover, there is consistency between their sense of a moral obligation to be honest and the tendency to overestimate the chance of being caught". In short, and even though there appear to be differences linked to other elements of the context (country, tax rates etc.), it seems clear that the probability and severity of punishment are *not* the primary drivers of tax compliance – but rather, that the moral values of taxpayers, and their views on the legitimacy of the tax system and its rules, are the fundamental drivers, to which inspections and enforcement only come as an addition[363].

Evidence from research on citizens, police and courts

Several decades of research on criminal matters, and on interactions between citizens and authorities (police and courts in particular), paint a similar picture to what we have seen for tax. Of course, deterrence does play a role in fostering compliance (i.e. deterring crime), but it tends to have an effect that is limited (or even very limited), except if considerable resources are expended so at to make the probability of detection really high. On the other hand, process-based factors appear to play a crucial role in determining sustained attitudes in respect with laws and regulations, and with public authorities.

---

[360] Full quote: "In 1992, Fischer, Wartick and Mark reviewed a bulk of studies directed at learning more about the relationship between probability of detection and compliance behaviour. It appears that the reviewed studies, which employed different methods, generally point in the same direction and strengthen the confidence that increasing the probability of detection will result in less non-compliant behaviour. However, the effect is, if anything, very small. Similarly, while the effect of fines is significant in many studies, their impact on tax compliance in general is small, if not negligible (Andreoni, Erard and Feinstein, 1998)"

[361] Quoted by Kirchler from Fjeldstad and Semboja (2001) - study on tax behaviour in Tanzania.

[362] On tax morale in Switzerland by Torgler (2005).

[363] Quoting one last time from Kirchler (2007): "Based on the rather small effects of variables considered in the neoclassical economic approach (i.e., audit probability, fines, marginal tax rate and income), several studies conclude that it is important to consider also citizens' acceptance of political and administrative actions and attitudinal, moral and justice issues as they are central to psychological and sociological approaches (Lind and Tyler, 1988 ; Pommerehne and Frey, 1992 ; Pommerehne and Weck-Hannemann, 1992 ; Tyler and Lind, 1992 ; Weck-Hannemann and Pommerehne, 1989)."

Tom Tyler, summarizes deterrence's impact and limitations as follows (2003): "studies of deterrence (…) point to factors that limit the likely effectiveness of deterrence models. Perhaps the key factor limiting the value of deterrence strategies is the consistent finding that deterrence effects, when found, are small in magnitude. (…) A further possible limitation of deterrence strategies is that, while deterrence effects can potentially be influenced by estimates either of the certainty of punishment or its severity, studies suggest that both factors are not equally effective. Unfortunately from a policy perspective, certainty more strongly influences people's behavior than severity, and certainty is the more difficult to change. (…)To influence people's behavior, risk estimates need to be high enough to exceed some threshold of psychological meaningfulness." This means that, in practice, deterrence is impossible to achieve in most cases: the resources required would be far too high (in a world of limited resources, society cannot commit enough resources to deterring violations in each and every regulatory field), and the intrusion on privacy and limitations of individual freedoms would be far too high. Tyler cites murder as a key example: on this topic, society has allocated enough resources that indeed there is a real deterrence effect – but achieving similar intensity of enforcement in all other fields is impossible. In addition, deterrence approaches "are not self-sustaining and require the maintenance of institutions and authorities that can keep the probability of detection for wrongdoing at a sufficiently high level to motivate the public."

By contrast, process-based approaches aim at increasing the legitimacy of rules and authorities by improving the level of fairness as perceived by citizens. The focus is not primarily on "distributive justice" (i.e. having *outcomes* that are deemed fair) – although this also has been found to have a significant impact on compliance, it is significantly less strong than the process effect, and in addition it is in practice impossible to reach decisions that would satisfy everyone. Rather, the emphasis is on "procedural justice". In the words of Tyler (2003), who has been one of the key proponents of this approach for several decades: "The procedural justice model involves two stages. [First,] public behavior is rooted in evaluations of the legitimacy of the police and courts. (…) In other words, people cooperate with the police and courts in their everyday live when they view those authorities as legitimate and entitled to be obeyed. [Second,] the antecedents of legitimacy. The procedural justice argument is that process-based assessments are the key antecedent of legitimacy (…). In this analysis, four indicators – summary judgments of procedural justice, inferences of motive-based trust, judgments about the fairness of decision making, and judgments about the fairness of interpersonal treatment-are treated as indices of an overall assessment of procedural justice in the exercise of authority." Crucially, research has shown that the effect of procedural justice is *significantly stronger* than that of deterrence.

The procedural justice effects are found in many fields and settings (mediation decisions Lind et al. 1993, dismissal from employment Lind et al. 2000 etc.). What also matters is that procedural justice, and the legitimacy it fosters, are long-term drivers of compliance, and largely self-sustaining (at least they do not require an *increase* in resources – but a change in behaviours and approaches). The changes involved in how authority is exercised are, however, significant compared to what is the practice in many cases. Quoting Tyler (2003) again, the key conditions needed to achieve a procedural justice effect are: "that decision making is viewed as being neutral, consistent, rule-based, and without bias; that people are treated with dignity and respect and their rights are acknowledged; and that they have an opportunity to participate in the situation by explaining their perspective and indicating their views about how problems should be resolved."

Refining the models: *responsive regulation, smart regulation*, and their challenges

Responsive regulation: the original vision

Confronted with the limitations and contradictions of simple models of understanding and fostering compliance, scholars (and practitioners) have been developing more complex models, attempting to *combine* several approaches in a coherent framework. The first, and arguably the most famous, is the *responsive regulation* model that was formulated in 1992 by Ayres and Braithwaite (relying on earlier work, and later further developed by Braithwaite and others, in particular Grabosky). The fundamental idea of responsive regulation is that different approaches are needed (and warranted) for different businesses, that these different approaches need to be seen as part of a *pyramid* of *escalating severity*, and that the regulators need to be *responsive*, i.e. change approaches as business behaviours change. In addition, they argue that the overall "enforcement pyramid" needs to be publicized so that regulated entities know exactly what to expect, and thus have an additional incentive to comply, so as to remain at the "bottom of the pyramid" (Cf. Ayres and Braithwaite 1992, pp. 35-41).

The foundations for this model, precisely, combine different compliance drivers, and recognize that different businesses (and different employees within them) may be at different points in the pyramid (corresponding to different drivers being strongest), and that their position may change over time (in particular in reaction to regulatory enforcement actions). The bottom of the pyramid corresponds to pure persuasion, while successive moves *up* the pyramid correspond to increasingly strong deterrence (and, ultimately, incapacitation) (p. 35). Of course, the precise list of actions will depend on the context, and in particular on the legal tools available to the agency.

The *shape* of the pyramid is meant to convey the idea that "most regulatory action occurs at the base of the pyramid where attempts are initially made to coax compliance by persuasion" (*ibid.*). The same pyramid model is suggested for *enforcement strategies* – meaning that, at a strategic level, governments and regulatory agencies should tailor (and communicate) their strategies in the same way. The specific regulatory instruments included in this pyramid can vary (the authors present an example ranging from self-regulation to "command regulation with non-discretionary enforcement" at the top, on p. 39) – but the general benefit is that "clear communication in advance of willingness by the state to escalate up the pyramid gives incentives to both the industry and regulatory agents to make regulation work at lower levels of interventionism", in the hope of avoiding the "cost of increasingly inflexible and adversarial regulation" (pp. 38-39).

In order to work effectively, responsive regulation requires that regulatory agencies have at their disposal a broad range of potential responses (including varied sanctions of increasing severity), that allow them to have an enforcement approach that can be as "finely graded" as possible. By contrast, if an enforcement agency has only very severe sanctions available, the threat to "cooperate or else" will not be credible because regulated entities will know that this (exceedingly drastic) sanction will usually *not* be used. When the different sanctions available do not fit well with the range of severity of possible offences, there will be situations where there is "no politically acceptable way of punishing these offences" (pp. 36-37)[364]. We would add, writing from experience in very different jurisdictions, that this last point is true for democratic polities, and even (within these) for polities with a strong voice for businesses. There are a number of situations where such exceedingly severe sanctions *will* be used, and where the effect will be that not only will violations be deterred, but

[364] There are ways to introduce "nuances" in practice with what appears at first to be a limited "response kit". For instance, Hawkins (2002) shows how British HSE inspectors developed rather sophisticated techniques of persuasion to address the limitations of their available range of responses – with formal enforcement including only improvement notice, prohibition notice and prosecution. Tilyndite (2012) argues in fact that their use of notices has been so effective as to make the introduction of administrative penalties rather unattractive.

legitimate investment and economic activity, with serious consequences for growth and employment. Ayres and Braithwaite also suggest the idea of what they call a "benign big gun" (pp. 40-41), where enforcement agencies have sanctions reaching *very high* (until full incapacitation), so that (using the pyramid approach) they can use the threat of this power to, in fact, push *most* regulatory interactions to the bottom end of the pyramid. The combination of responsiveness, gradation and very high "top of the pyramid" would thus function optimally.

The responsive regulation approach incorporates earlier findings and ideas on "tit-for-tat" (in particular Scholz's work – cf. pp. 20-23), but goes significantly further. Indeed, "tit-for-tat" is premised on assumptions of rational behaviour, and can be formulated through a game-theoretical analysis (cf. pp. 21, 60-81). Ayres and Braithwaite's vision, by contrast, incorporates a complex view of compliance motives (what they call "mixed motives" – cf. pp. 22-35), and thus modifies the "tit-for-tat" vision into the "compliance pyramid". The pyramid acknowledges that different motives work (to different amounts) for different people (and in different situations), and seeks to rely as much as possible on voluntary (values-based) compliance, while keeping deterrence (calculation-based) in the background. Being particularly open about the potential for very high escalation, but mostly *not* using sanctions, i.e. "speaking softly and carrying a big stick" is the core of their approach. But the vision they lay out in their 1992 book (as distinct from the many summaries produced later on by other scholars) has many other aspects. In particular, it envisions a strong reliance on "tri-partism" (cf. pp. 54-100), whereby the role of "public interest groups" (trade unions, NGOs and civil society organizations) would come into play to avoid regulatory capture, and ensure more optimal outcomes than a simple two-way relationship would (cf. pp. 86-97). They also discuss at length the potential for "enforced self-regulation", and the different ways in which it can be structured, as a potential application of the "pyramid" approach (pp. 101-132). While these are very interesting directions for reflexion, and they are *connected* to our area of research, their relevance to our concerns is at this point marginal, and we will not discuss them further[365].

From "smart regulation" to "really responsive regulation"

Already, through their vision of "tri-partism", Ayres and Braithwaite started to consider the importance of *other actors* in the question of compliance and of public welfare outcomes. Gunningham and Grabosky developed this further into an approach they called "smart regulation" (1999), a term which quickly became used in a confusing variety of ways. Their understanding of the notion was "an emerging form of regulatory pluralism that embraces flexible, imaginative and innovative forms of social control which seek to harness not just governments but also businesses and third parties" (Gunningham 2010, p. 131). Their fundamental insight is that there are many influences that shape business behaviour, far beyond regulation (and simple cost-benefit calculations related to narrowly-defined compliance), such as "international standards", "trading partners and the supply chain", "financial markets", "peer pressure", "internal (…) culture" and "civil society" (*ibid.*). This is an approach that we see as very relevant, and indeed we will try and show in our examples from the practice (in the third part) that "risk-based" inspection systems tend to also try and leverage all or at least several of these different factors. Nonetheless, we will not discuss these in depth, as our focus in this research is specifically on the regulatory enforcement aspect.

---

[365] We will just note three more things about Ayres and Braithwaite's work. First, the "tri-partism" vision, while it is clearly very context-related (i.e. rooted in Australian conditions) is very interesting – and has clearly been influential in further research (see next paragraph on *smart regulation*). Second, "enforced self-regulation" can be linked to other models such as third-party conformity assessment in product-market regulations, or to the vision of "ethical regulation" by Hodges (2015) etc. It definitely warrants further research. Finally, the authors also call attention at the opening of the book to the relevance of their research not only to OECD countries but, for instance, to post-Soviet countries (p. 7). We very much agree, in spite of all the practical difficulties, as we will further develop in the third part of this research.

A number of criticisms, or remarks, have been done concerning the original responsive regulation model (and Braithwaite himself has introduced a number of additions to it). To our mind (and contrary to the way in which these remarks were sometimes done, which suggested that something was fundamentally amiss in the original design), they do not reflect any essential flaw in the design of responsive regulation, but rather point to specific points in detailed implementation, which could not possibly be all addressed in the original work, which was rather short and conceptual. One of the most important points is the difficulty to "ratchet down" enforcement, "rebuilding trust" after an escalation (Gunningham 2010, p. 127 – quoting Haines). Other points relate to the fact that, in many situations, the ideal "pyramid" model will not be (or not be fully) applicable. For instance, interactions may be too rare (or too rarely repeated), or several regulators may be involved (with different approaches) (*ibid.,* pp. 128-130). The second problem (several overlapping regulators with inconsistent approaches) is indeed a frequent problem – in our view, however, it does not suggest anything wrong with the responsive regulation model, but with the institutional setup (as there are many other reasons why overlapping, uncoordinated regulators covering the same issue are *not* an optimal setup)[366]. The first problem, however, is quite rare in fact in our experience. While indeed some very small agencies, or agencies covering very specific issues (such as fisheries inspections used as example by Baldwin and Black 2008), may have very rare interactions with regulated entities, it remains that the inspectorates which "matter" in the experience of businesses (and, generally, in the perceptions of the public) have typically rather large staffing levels, and relatively frequent interactions with businesses. In addition, if and when there are violations or problems found, re-inspections are relatively frequent, and thus the problem of "too rare repeat interactions" is not, in our experience, a very serious challenge to the responsive regulation model.

Another view, which is more relevant in practice, is that in many instances it would be sub-optimal to rely *only* on the pyramid, for a variety of reasons. The first is that interactions, while not being necessarily so rare that the pyramid is inapplicable, can be relatively infrequent (e.g. for small, low-risk businesses), and thus the pyramid is a less-optimal approach than *segmentation,* whereby regulators select "the most appropriate regulatory tool from a variety of options" for a given target group or entity (Gunningham 2010, p. 130). Likewise, there may be case where interactions' frequency is not the issue, but where "the classification of regulated enterprises into one of a variety of motivational postures" is "relatively straightforward". In such cases, a "target-analytic" approach can be more efficient than a "tit-for-tat" one (*ibid.*, p. 128). Again, we see here nothing actually *contradicting* what Ayres and Braithwaite outlined, particularly if one considers that they specifically suggested having a pyramid of *enforcement strategies* and not only one of enforcement *responses* to a given entity. Such selection of tools based on profiling can perfectly fit the perspective of an enforcement strategies pyramid.

Developing a number of these criticisms, concerns and additions, Baldwin and Black have written two papers on "really responsive regulation" and "really responsive risk-based regulation" (2008 and 2010, respectively). These are important contributions, and try and integrate a number of different strands of scholarship and practice – responsive regulation, risk-based regulation, Sparrow's "regulatory craft" approach, and close consideration of practical challenges of regulatory agencies. Here again, we would argue that the way the authors present several points as criticisms or contradictions of the original responsive regulation framework somewhat overstates the real differences – which are more about nuances, practical applications, and consideration of implementation challenges. Nonetheless, they make a number of very important points. First, they rightly point out that the pyramid needs to be combined with a risk-proportionate response: "

[366] See Blanc (2012) pp. 22-25

in some circumstances step by step escalation up the pyramid may not be appropriate. For example, where potentially catastrophic risks are being controlled it may not be feasible to enforce by escalating up the layers of the pyramid and the appropriate reaction may be immediate resort to the higher levels" (p. 6)[367]. Second, they emphasize the fact that "tit-for-tat" may be wasteful when it is clear which approach is appropriate (or not) for a group of regulatees (p. 7 - again, this is a point which the pyramid of enforcement approaches would, in principle, cover). Third, they point out the problem of regulatory regimes where inspection and enforcement activities are "spread across different regulators with respect to similar activities or regulations" (p. 8) – a point which, as we noted above, is very important in practice, but says little about the model, and more about institutional problems that require a solution (because incoherent and inconsistent enforcement would be a problem with or without responsive regulation).

Baldwin and Black go on to make a certain number of recommendations to achieve "really responsive" regulation. These are sound recommendations, that mostly relate to the attention to *implementation*. They include paying attention to "the constraints and opportunities that are presented by the institutional environments within which the relevant regulators act" (p. 19) and attention to the "logics of different regulatory strategies and tools" (which involve "different understandings of the nature of behaviour or of an institutional environment, and in turn have different preconditions for effectiveness" – p. 20). They stress the crucial importance of "responsiveness to the regime's own performance and effects", and thus of developing adequate tools for "performance evaluation *and modification*" (emphasis ours – p. 21)[368]. On this basis, they develop a set of key questions covering the five challenges of "detection, response development, enforcement, assessment and modification" (p. 26). While this is a very interesting grid to assess regulatory responses, much of it goes beyond the scope of this research, were we really focus on the enforcement phase (and, to some extent, on response development). Detection problems (cf. p. 30-31) are also very relevant for inspections issues, and we will to some extent discuss them in the third part. We would, however, suggest that they are *in many cases* somewhat less acute than Baldwin and Black suggest. First, because the example they use (fisheries regulation) is particularly extreme, and detection is far easier in many of the more "common" regulatory functions and regulated sectors. Second, because if detection issues are really so considerable for a given type of inspections that they make it essentially impossible or ineffective, then this issue should be considered at an earlier stage of regulatory design, i.e. when identifying the problem and coming up with a regulatory solution. If inspections *cannot* realistically work, then maybe they were never the right tool in the first place[369].

In conclusion, one could say that the responsive regulation model, with a number of additions and nuances, has given a solid basis for further theoretical and practical developments, by formulating a coherent framework which allows differentiated approaches based on context, target group, interaction history etc. "Smart Regulation" and successive contributions have brought more attention to multiple stakeholders and tools, and to implementation challenges. "Meta Regulation" (Cf. Gunningham 2010 pp. 135-139) has suggested to develop Ayres and Braithwaite's vision of "enforced self-regulation", looking at systems put in place by firms themselves, and verifying their effectiveness. All of these additional models and contributions

[367] Nothing, in the original *Responsive Regulation* model, suggests that *no other factors* should be taken into account – and, to us and to many practitioners, it is clear that they should be combined (as they are in OECD 2014) with risk proportionality. This articulation is, however, missing from the original model which, as we have seen, was rather short and conceptual in most areas.

[368] While we will return several times to the question of *measuring* effectiveness, discussing the challenges of *transforming* practices and institutions on the basis of performance evalutions would go beyond the scope of this research.

[369] For radically different approaches of fisheries regulation see e.g. Eythórsson 1996 or Runolfsson 1997. Measuring the performance of different approaches, which Baldwin and Black see as a very problematic, could arguably be done through looking at fish stocks evolution rather than at compliance. It is worth considering, in cases that appear extremely problematic, whether the reliance on command-and-control regulation, enforcement and compliance is possibly not the right approach.

consolidate the view that it is most effective to rely on a combination of tools and approaches, based on the specifics of the regulated entities and the regulations being enforced, on the context, and on prior history.

The challenges raised by Baldwin and Black (2008, 2010) also remind us of the importance of *first* thinking through whether command-and-control regulation, and subsequent enforcement efforts, have any chance of success at solving the problem at hand, and are likely to be an effective and efficient response – regardless of the specific inspection approach taken. In many cases, the answer may simply be negative, and other policy interventions will be more adequate[370] (see Ogus 1994 for an overview of other regulatory tools, and all the literature on Regulatory Impact Assessment for discussions of other policy options). Using methods that come from beyond the narrow "regulatory" field may also bring major benefits from this perspective. Increasingly, the "causal pathway" methodology[371] is being used in the regulatory and enforcement area, to determine what are the mechanisms that cause the unintended effects (increased risk, decrease in public welfare) that regulation is meant to address. The models allow to consider whether regulation is really likely to be useful and, if so, which intervention mechanisms may be most helpful (see BRDO 2013 for a practical application of this methodology).

Diverging data, diverging conclusions

In spite of some meta-studies (like Kirchler 2007) seemingly indicating some more-or-less clear trends, the data on compliance effects of different approaches is, in fact, disputed. As Simpson and Rorie (2011) put it, there are "several general traditions in this regard, each with its own logic and empirical base" (p. 59), which is a way to say that different streams of research seem to come up with data that cannot fully be reconciled. For instance, while we have quoted above Tyler at length, and his findings on the strength of procedural justice effects (confirmed by a number of other scholars), some research seems (at least at first glance) to question his confidence (in the possibility to found compliance primarily on "procedural justice" ("studies find that the primary factor shaping legitimacy, morality and rule adherence is the procedural justice that employees experience in their workplace", 2011 p. 78). Similarly, it is not certain that Hodges's (2015) confident assertion that "public enforcement based on a policy of deterrence does not "in fact [have a] significant deterrent effect" (p. 26).

In the interest of presenting the evidence in a clearer way, we have of course somewhat over-simplified the different perspectives, and it behoves the topic's complexity to add some important nuances. First, a distinction is often made between "general deterrence (premised on the notion that punishment of one enterprise will discourage others from engaging in similar proscribed conduct) and specific deterrence (premised on the notion that an enterprise that has experienced previous legal sanctions will be more inclined to make efforts to avoid future penalties)" (Gunningham 2010, p. 122)[372]. In addition, evidence "shows that regulated business firms' *perceptions* of legal risk (primarily of prosecution) play a far more important role (…) than the objective likelihood of legal sanctions" in determining general deterrence's effectiveness (*ibid.*). Thus, there is maybe not a sharp distinction to be drawn between "rational" motivations (calculations) and other drivers – since even so-called "rational" deterrence estimates are based on perceptions rather than on objective data, and perceptions appear to be strongly interrelated with values-based thought processes.

[370] A point very similar to that made by Ashworth (2000) on the excessive use of criminal law for problems where it is inadequate.
[371] Which is also widely use in a number of domains (political science, ecology, epidemiology etc.) – see a theoretical summary on Cornell Evaluation Centre website: https://core.human.cornell.edu/research/systems/theory/causalpathways.cfm
[372] The "evidence of a link between past penalty and improved future performance is stronger", suggesting that *specific deterrence* is more powerful than general one – but research shows also that "action falling short of prosecution" can achieve substantial effects, i.e. that it is more the *warning effect* rather than the punishment which matters (Gunningham 2010, p. 124).

Indeed, research suggests that many respondents struggle to "disentangle normative from instrumental motivations" (*ibid.,* p. 123).

The effect of deterrence may also vary over time, or be complex rather than linear. As Gunningham writes, "it is plausible (…) that the deterrent impact of tough enforcement may be weaker today, than it was in past decades, at least in industries that have been subject to substantial regulation for a considerable period and/or are reputation sensitive" (*ibid.*). This is clearly hypothetical, and *may* be true in some cases – while there is also some evidence that, possibly in other regulatory fields and/or countries, *new* regulation that is primarily implemented through "tough" enforcement tends to fail, and more "persuasion-grounded" efforts fare better. There also is evidence of weak deterrence effects in completely different contexts from Gunningham's. This all suggests that there may be specific contexts where the effects are stronger or weaker, which cannot however just be explained by one variable, but rather by a combination of many factors.

One study, many findings, complex interpretation

In a study of Chinese farmers, the findings of which were published in a series of papers in 2015, Yan, van Rooij and van der Heijden attempted to observe directly the actual level of compliance, and to assess the strength of different drivers through interviews. They took a very comprehensive and "non-partisan" view of compliance drivers, looking at the whole range: ability to comply (physical/economic capacity, legal knowledge), deterrence, procedural justice, prevailing social norms, and internalised moral duties (2015 b pp. 2-3). Their findings, though founded on a study of only a bit over 100 farmers, are highly interesting – and require careful interpretation.

In a first paper (2015 a), the authors simply crossed the different types of behaviour (compliant/non-compliant for three different norms on pesticides) with the different drivers of compliance (ranked as positive or negative), and examined correlations. They had several conclusions: first, that deterrence was overall limited in effectiveness, not because of an absence of correlation between probability of detection and compliance, but because high probability of detection seemed to be closely correlated with a high level of other (voluntary compliance) factors. In other words, the farmers most frequently inspected (the large farmers, as the authors found) were also those that were anyway the most likely to comply even without inspections and enforcement (pp. 7-8). Second, that "apart from deterrence, operational costs and benefits, personal norms, social norms, and, less clearly, legal knowledge all play a role in compliance" (and that this role is significant) (p. 13). By contrast, the authors found no "clear relationships between the general duty to obey the law, procedural justice and compliance", leading them to add that "these variables are not crucial aspects of voluntary compliance, and thus enforcement does not have to take them into account" (p. 11).

It is worth, however, pointing out a few points from the authors' *data*, which may support different interpretations. First, across the board, compliance (and apparent responsiveness to "drivers") is strongest for the type of regulation that is the most directly understandable and, arguably, has the greatest safety effect: the prohibition of some hazardous types of pesticides. Rules on disposal and time interval before marketing are far less well respected (a point the authors note, but do not necessarily pay enough attention to). Second, procedural justice is overall quite consistently low: most respondents have a feeling of negative procedural justice. It may simply be that, in a context where interactions with authorities are nearly uniformly marked by "rough handling" and top-down commands, farmers simply fail to register the very few exceptions as being significant. This does not *ipso facto* mean that they would not respond to a sustained experience of a different approach, or that this type of authoritarian behaviour has no negative effect (e.g. possibly on the overall respect for laws etc.). Finally, the authors' conclusion that the targeting used by Chinese inspectors (who primarily inspect the larger farms, which are found to be the ones most likely to be voluntarily compliant) is wrong can also be disputed. Their view is that this results in deterrence failing to have an impact on those

most likely to be non-compliant, and write "as a matter of principle, enforcement should be targeted especially at those types of farmers and those types of rules for which voluntary compliance is less likely" (p. 13). We will come back further below to the question of defining "risk" and of what targeting makes more sense – but from a practical perspective, targeting the highest-impact farmers is far from being irrational. Furthermore, the authors have no way to be sure that (a) frequent inspections of larger farmers have not played a role in them understanding rules better, and being generally more supportive of voluntary compliance and (b) inspections and enforcement would be the most appropriate tool to target smaller farmers and increase their knowledge and voluntary compliance (in fact, it is possible that inspections would increase costs, through time lost and sanctions, and thus further decrease their financial capacity to comply – for instance).

Their second article drawn from the same data (2015 b), but with a different analytical methodology ("crisp set Qualitative Comparative Analysis" – csQCA)[373], yields conclusions that are somewhat different, and very interesting. The most striking result is the absence of equivalence between the conditions of compliance, and that of non-compliance: "our data also point to a non-symmetrical relation between the deterrent effect of sanctions and compliance. The analyses of necessary conditions pointed out that (experienced) deterrence is not a necessary condition for compliance, but that the absence of (experienced) deterrence is a necessary condition for non-compliance (…). Our data further indicate that deterrence (as part of a set of compliance conditions) does play a marginal role in affecting compliance (in one path for one compliance behaviour), the absence of deterrence does, however, play a considerable role in affecting non-compliance (in five out of six paths for both non-compliance behaviours). This finding challenges our thinking about the assumed compliance–non-compliance dichotomy in the literature—it indicates that compliance is not necessarily the inverse of non-compliance" (p. 14). This suggests that, while deterrence may indeed be of little importance for active compliance (voluntary compliance drivers being sufficient), in order to "drift" into non-compliance, the absence of a significant deterrence effect is an important "trigger". The authors interpret it thus: "deciding to comply is not the same as (also) deciding 'not to violate'—if we decide to stick to speed limits, we likely do not (also) decide to not hit the pedal to the metal (building on insights from behavioural economics, cf., Kahneman 2011). What we observe is not a reassurance and reminder function for compliant decision making from deterrence, but the lack of deterrence as a reminder and reassurance that violation goes unnoticed or unpunished (…). This reasoning is in line with Ian Ayres and John Braithwaite's (1992) responsive regulation model, which assumes that most compliance will occur without active deterrence" (p. 15). Looking more closely at the conditions for non-compliance, one finds "the absence of a deterrent effect of sanctions, a non-positive cost-benefit analysis, the absence of (experienced) descriptive social norms, and the absence of (an experience of) procedural justice" but with "relatively low coverage scores" (i.e. weaker effect) for deterrence and procedural justice. The conditions for compliance, by contrast, include (different combinations in different compliance paths) "law as a source of moral authority", "descriptive social norms to comply", "positive cost benefit analysis", "legal knowledge", "capacity to comply", and "general duty to obey" – with "law as a source of moral authority" being present in every path.

Complex processes, nuanced conclusions

What all the evidence summarized above suggests is first that interpreting results and compliance processes finely may be vital: deterrence may well be superfluous for the majority (of voluntary compliers), but would

[373] As the authors explain: "QCA differs from other methods in its focus. 'The key issue [for QCA] is not which variable is the strongest (i.e., has the biggest net effect) but how different conditions combine and whether there is only one combination or several different combinations of conditions (causal recipes) of generating the same outcome' (…). QCA is grounded in set theory, a branch of mathematical logic that allows the study, in detail, of how causal conditions contribute to a particular outcome. A particular strength of QCA is that it can be applied to arrive at evidence-based typologies" (p. 5)

be needed (and need to be targeted) for those who are "on the brink". Second, the fundamental drivers appear to be social and personal norms and ethics: accepting the law as source of moral authority, following social norms, feature as the strongest drivers. Third, capacity to comply is also crucial: financial and physical, as well as (to a lesser extent) legal knowledge – but also cost-benefit analysis. In other words, norms that are realistic given prevailing conditions, well explained and communicated, and tailored so as to be economically viable, stand the best chances of success (a relatively unsurprising finding, one may add, but still an important one). Finally, the variables that correspond most closely to traditional "enforcement" (both deterrence and procedural justice) appear, in fact, the weakest (and this second paper puts procedural justice alongside deterrence, somewhat nuancing the conclusions of the first one). What fundamentally matters is whether citizens (including those that work in businesses) adhere to norms that make them comply – not whether they are checked frequently, and how. Inspections and enforcement work at the margin. Precisely because they work at the margin, we would add, means that they should use each and every tool at their disposal (including procedural justice) to be more effective – because affecting social and individual norms is, at best, a long-term undertaking[374]. In fact, all these findings (including the importance of information) all match what many practitioners know and do, at least in what one could call "smart inspections" regimes, as we will discuss in the third section.

Thus, the overall importance of "enforcement" aspects for compliance may be relatively weak compared to deeper, longer-term factors – and in addition the respective strength of different aspects and factors is, as we have seen, not so easy to ascertain, and/or varies according to circumstances. In addition, different approaches carry some trade-offs, that are in evidence in a number of studies.

Charting a course in spite of uncertainties

There may be some ways to move forward, and try and make sense of inspections and enforcement methods, in spite of these uncertainties. This requires first to understand how *context* may cause differences in results, then to acknowledge the *limitations* in methods and findings – and finally to suggest *alternative* sources of evidence.

Context and typologies

The models of compliance we have outlined above seem to a significant extent to be contradictory, and conflicting research findings do not lead to an easy way to decide upon their validity or to reconcile them (even though some models do seem to be more strongly validate than others). This creates difficulties for our research object, since the question of compliance drivers is essential in order to provide a foundation for the choice between different enforcement approaches. A way to make sense of these contradictions, and to end up with a model that somewhat reconciles different drivers and perspectives, is to consider *context* and *typologies.* Context, because one of the reasons different models appear to be validated (or invalidated) by different studies may be that some compliance drivers are stronger (or weaker) depending on the broader circumstances where they apply. Typology, because it may also be that different drivers apply to varying

[374] Simpson *et al.* (2013) reach a somewhat similar conclusion in their assessment of crime-control strategies for corporate environmental crime: "First, both informal sanctions and command-and-control strategies lower the likelihood of corporate crime. The risk of corporate offending increases when there is not a credible legal threat or when one's duty to behave ethically is not reinforced by colleagues or through fear of informal sanctions. Second, the deterrent capacity of these control mechanisms does not negate certain corporate or individual risk factors, which remain significantly associated with noncompliance. This suggests that current policy levers do not fully mitigate offending risks and may indicate that a one-size-fits-all policy is shortsighted." (p. 267)

extents and with varying strengths to different types of people (or of groups of people). This is what we will now attempt to consider.

As Kirchler (2006) puts it, the research on tax compliance has come into its own "as a research area within economics and economic psychology", and studies have "considerably increased" over a few decades, but there is serious concern "since the results obtained in different studies are heterogeneous" (p. 1). The way he suggests to make sense of these (apparent, at least) contradictions is that "some heterogeneity in results can be reconciled by considering the relationship between the authorities and the taxpayer". In other words, the drivers of compliance may be different (or at least have different relative salience) in a "*cops and robbers* climate" and in a "*service for clients* climate" (*ibid.*). In a climate of distrust, the primary driver will be deterrence, based on rational calculations – so compliance will occur only if detection and sanctions are really credible. In a climate of trust, social representations, norms and fairness perceptions will be the main drivers (pp. 1-3). From this, Kirchler proposes a model of three-dimensional representation of compliance, whereby the dimensions are compliance, power of authorities and trust in authorities (pp. 8-9). An enforced compliance approach will tend to succeed only if it can maximize power, a voluntary compliance one if it can maximize trust – and there is a significant amount of trade-off between the two, because "sharp undifferentiated control" and "severe punishment" tend to result in a sharp reduction of taxpayers' willingness to comply voluntarily (p. 6). In other words, different findings e.g. regarding the effectiveness of deterrence activities, or of procedural justice aspects, may reflect at least in part different contexts[375], in which ongoing relationships between administration and taxpayers have shaped certain attitudes and expectations[376].

Clearly, such aggregate differences also cover different types of compliance profiles within a given society – linked both to individual and social differences, resulting e.g. in differences in perceptions of fairness (p. 16). This results in different profiles, "motivational postures" as V. and J. Braithwaite have called them – ranging from "commitment" through "capitulation", "game playing" and "disengagement" through "resistance" (p. 17). As Kirchler points out, such findings strongly support the "responsive regulation" approach, which allows to tailor the type of response to the type of regulated person or entity, and to minimize the use of deterrent enforcement (thus minimizing the negative responses which weaken voluntary compliance).

Another possible typology, proposed by Elffers and Hessing (1997) and taken up by Voermans (2014) distinguishes "conformist compliers" ("those who comply with rules only because they fear punishment"), "identifiers" ("comply with rules because they want to belong to a social group for which compliance is the norm") and "internalisers" ("who comply with rules because they have made these rules part of their own world view")[377]. These different types require different responses: sanctions are superfluous for the third group, and have an indirect effect on the second (serves to "maintain the social norm" by showing that infringements are punished). For the first group, sanctions *can* be effective but need to be "*certain*, *quick* and *severe*", which is difficult to achieve – and may be counter-productive, as we have seen that systems which deploy an excessively harsh deterrence approach tend to weaken voluntary compliance.

---

[375] This perspective is very important because it also suggests that it may be difficult to move from one approach (*cops and robbers*) to the other (trust-based). This is certainly what our experience in post-Soviet states suggests. Years of outright hostility from inspectors towards businesses have yielded a situation where gaming the system is the norm, trust is non-existent, and moving to a better situation is extremely difficult. The situation in these countries also strongly validates Kirchler's concerns about how to "control the controllers" in systems based on distrust. Indeed, inspectors in such countries tend to abuse their powers routinely – and this is a risk in any system where regulated subjects are seen as suspects, and controllers vested with very strong powers (and few checks and balances).

[376] Taking also into account, as Kirchler emphasizes, that "perceptions" and "representations" (what people think about the authorities, the tax system etc.) are in practice more important than "what actually is" (p. 13) – and that what is fundamental is the overall "aggregate" of "knowledge, attitudes, norms, perceived opportunity, fairness considerations and motivational postures", i.e. "tax morale" (p. 17).

[377] Voermans (2014) p. 57.

Gunningham (2010) lends additional support to such views by indicating the complex, intertwined workings of "normative" and "instrumental motivations": many business operators "wrestled with the temptation to backslide when legally mandated improvements proved very expensive" and "many acknowledged that, in the absence of regulation, it is questionable whether their firms' current good intentions would continue indefinitely – not only because their own motivation might decline, but because they resented others 'getting away with it'" (p. 123)[378]. He also warns against the other downside risk, which may materialize when excessively harsh and "across the board" deterrence approaches are used – fostering a "culture of regulatory resistance", and "being counter-productive as regards corporate leaders who respond badly to an adversorial approach" (even as it may be "effective when applied to the recalcitrant and perhaps to reluctant compliers") (p. 124).

It is difficult, of course, to estimate how many businesses or operators may belong to each category. Bardach and Kagan (1982), suggested a rule of thumb of 20% of "bad apples" and 80% of "good apples" (p. 65), founded on several studies and testimonies. In particular, they reported that a "study of housing code enforcement in New York City found that 65 percent of recorded violations were attributed to 12 percent of all multiple-dwelling buildings" (*ibid.*). They also quoted the reflection of the WWII head of the Office of Price Administration in the US, Chester Bowles, that "20 percent of the regulated population would automatically comply (…) simply because it is the law of the land, 5 percent would attempt to evade it, and the remaining 75 percent would go along with it as long as they thought the 5 percent would be caught and punished" (pp. 65-66). The authors' conclusion is that "the absolute and relative proportion of good apples is large, almost certainly constituting a sizable majority (…) with respect to most regulatory domains". They note that "the absolute number of bad apples is also large" but that "ready recourse to coercion" and "uniform, specific regulatory prescriptions" that may be necessary for "bad apples" can, when applied to "good apples", lead to a "considerable amount of unreasonableness" and unintended adverse consequences (p. 66).

Building on these different but concurring views, it is worth adding that these typologies need not be understood as categories in which businesses, or people, can be ascribed permanently. Depending on the circumstances, the type of rule being considered[379], the administration with which one is dealing, the same person may have very different behaviours, and could be categorized in one or the other group. This is even more true when considering a complex entity such as a business, where different workers and managers may be significantly different. If this perspective is correct, then indeed different findings may simply reflect different situations, and the "optimal" enforcement strategy would be one that seeks to combine all different drivers[380], with careful attention to the risks of negative interactions between them, e.g. of deterrence weakening voluntary compliance. "Responsive" and "smart" enforcement would appear to fit best with such a perspective.

Limitations of methods

Science in general is difficult and, by definition, provisory and uncertain (radically so if we take Popper's definition of science as being characterized by "falsifiability"). Social science and psychology are made even more difficult by the complexity of their objects, and the considerable difficulties involved in measurement.

---

[378] See also *ibid.* the risks when excessively "persuasion-based" enforcement strategies "degenerate into intolerable laxity" (p. 125).

[379] As we have seen above in Yan, van Rooij and van der Heijden (2015 a), compliance levels differed strongly for three different types of rules, and this could not be explained fully by differences in probability of detection. Illustrations of this point are easy to come by, and it is frequent to find that the same person will have different attitudes concerning different parts of the traffic rules, or between the tax code and the prohibition of theft and murder, for instance.

[380] A related perspective is that of Badarch and Kagan (1982), whose model is fundamentally deterrence-based, but who acknowledge the number of adverse, unintended consequences of "pure deterrence", and seek how effects could be achieved at lower costs, and with less adverse effects on compliance through deterrence-induced "resistance" (see pp. 96-97).

Kirchler (2006) points out many of these issues in relation with tax compliance: difficulty to measure evasion, different definitions of concepts, etc. He finds, in particular, significant problems with surveys, because of lack of correspondence between "respondents' self-reports of tax evasion and officially documented behaviour". He also points out the limitations of models, which inherently tend to reduce complex phenomena to a limited number of variables (p. 17-18). A recent major study looked at the "reproducibility of psychological science" and found that a large proportion of original findings could *not* be replicated (only 36% of replications resulted in a statistically significant effect that was similar to the original one – while in a number of other cases the results appeared to be somewhat similar, but not fully statistically significant etc.)[381]. As the authors emphasize, "how many of the effects have we established are true? Zero. And how many of the effects have we established are false? Zero. Is this a limitation of the project design? No. It is the reality of doing science, even if it is not appreciated in daily practice. Humans desire certainty, and science infrequently provides it. As much as we might wish it to be otherwise, a single study almost never provides definitive resolution for or against an effect and its explanation." Accepting and understanding these limitations is essential (and, we would add, not always understood, both by scholars and by those who use their results). This is a very complex field of research, and one that is only a few decades old. Studies generally have a number of limitations, including size. Thus, rather than expecting total certainty, we should draw from this wealth of findings a nuanced view, with more frequently convergent results suggesting that some effects may be stronger or more reliable than others[382].

Putting too much faith in the result of one or a few studies is one risk. Putting too much faith in explanatory theories and models is another one. As Ariel Rubinstein, one of the founding fathers of game theory, himself wrote: "there are those who believe that the goal of game theory is ultimately to provide a good prediction of behavior (…) I am not sure on what this vision is based". He adds, "then there are those who believe in the power of game theory to improve performance in real-life strategic interactions. I have never been persuaded that there is a solid foundation for this belief" (p. 634). He suggests, by contrast, that "the object of game theory is primarily to study the considerations used in decision making in interactive situations. It identifies patterns of reasoning and investigates their implications on decision making in strategic situations. According to this opinion, game theory does not have normative implications and its empirical significance is very limited. Game theory is viewed as a cousin of logic. Logic does not allow us to screen out true statements from false ones and does not help us distinguish right from wrong" (*ibid.*). This is an important reminder. Ayres and Braithwaite, in *Responsive Regulation* (1992), had an entire section (pp. 60-81) devoted to a game theory perspective of "tit-for-tat" and tri-partism. While they acknowledged the limitations of the model, Rubinstein's words should serve us to take such models *in general* with caution[383]. They can be useful as explanations of what the authors think, and of logical interactions, but putting too much confidence in their explanatory or predictive power is fraught with dangers. Once again, we are led to a posture that is one of modesty: no explanation or model is likely to have all the answers, and trying to combine different perspectives may be a safer and sounder approach.

---

[381] Open Science Collaboration (2015) – full text available at: http://www.sciencemag.org/content/349/6251/aac4716.full#corresp-1. Since the authors first screened the studies for which replication would be attempted, and only took those that cleared a number of quality hurdles, the percentage from *all* studies (taken at random) would be even lower.

[382] A good example of how scope and duration may produce interesting results that may not be visible otherwise is provided in Wittberg (2006). In the experiment he relates, the Swedish tax administration undertook a long-term campaign to strengthen "tax morale" through education, and regular (large scale) surveys to measure changes. The results appeared to be strongly positive – meaning that the fundamental social norms that are one of the foundations of compliance could be gradually changed (and that this could be measured). But such experiments have so far been very rare. Having more will require a substantial amount of time.

[383] One particular obvious weakness of game theory is its reliance on rationality – examples of real-life negotiations, such as those involving Greece and the Eurozone in the first half of 2015, show by contrast that actors are driven to a very large extent by ideological considerations and a variety of values. Had rational interest been the sole mover, and actors been entirely rational, the outcome of these negotiations would most likely have been very different, and come far earlier.

Caution about the strength of studies' results makes it more difficult to develop conclusions that may serve for further development of evidence-based policy making, and of evidence-based inspections and enforcement approaches in particular. If evidence is inconclusive, then deciding between competing views is hard. As we have seen, evidence is certainly not *fully* inconclusive. There appears to be many studies finding compatible or partly similar results, and ways to reconcile many of the apparently contradictory findings. Still, for all the reasons listed above, the foundation for evidence-based inspections, if it were limited to these different studies and models, would not appear to be as solid as we would wish it to be – and making any conclusions about, for instance, the effectiveness of risk-based inspections would be difficult.

What we will undertake to do in the third section of this research (after concluding this theoretical section by looking at research on risk and regulation) is, precisely, to look for such complementary, alternative evidence basis. As Kirchler (2006) and others have pointed out[384], the past decades have seen a number of inspections agencies (in the tax field as in others) transform their approach, moving for instance from a strict deterrence approach to a more "compliance-based" or "responsive" one. Some agencies have moved strongly in the direction of more risk-based inspections. Others, in the same or in nearby jurisdictions, have not done so.

For all these reasons, we believe there is value in looking for evidence in a different way – not only through focused studies (which yield more details and better attribution, but have a number of problem, as we have seen), and rather by comparing practices and aggregate outcomes of different inspection systems. A first possibility is to consider changes over time in the same jurisdiction - and a second one is to compare across different jurisdictions. The first approach is possible when there is a clear change (or at least a strong inflexion) in practices over a well-defined period of time, and when data on practices outcomes is available for the same period. The second is feasible when two or more jurisdictions, which are otherwise sufficiently similar, present sharply contrasted inspections practices, and have data of good quality and adequate for comparisons. We will see when considering concrete cases (in the third section) that it has proven more feasible to find examples of the second case than of the first – but at this point we will limit ourselves to a few clarifications of method.

When attempting to compare practices *and outcomes* between countries, or across time, the two parts of the comparison pose radically different problems. Outcomes, on the one hand, are relatively easy to define, at least for some of the major inspection functions: for instance reducing as much as possible occupational injuries and deaths, or deaths from food-borne diseases. The (considerable) problems stem from data reliability (often problematic, because of under-detection or under-reporting)[385], and even more strongly from attribution: how much can the level of a given indicator in a jurisdiction, and its evolutions, be attributed to inspections practices, which generally have only a minor influence compared to economic, technical, social and cultural factors? Practices, on the other hand, pose far less problems of attribution – even though they may be shaped by a number of factors, our main concern here are not the causes, but the practices themselves. Measuring them is, to some extent, difficult, because data on targeting is not public in most cases, and because the "qualitative" aspects can of course vary considerably between different inspectors, localities etc. We will see that *in practice* this potential difficulty can be to an extent overcome because the differences between different jurisdictions are in certain cases so considerable that, at least in first approximation, the underlying nuances can be discounted. Remains attribution as the main problem.

Here, there is certainly no perfect solution, and moving forward requires a set of assumptions. First, that while there are many factors influencing outcomes such as occupational safety and health, if all major known factors

[384] See in particular studies in Elffers, Verboon and Huisman (2006) and in Parker and Lehmann Nielsen (2011).
[385] And also, frequently, from different data definitions, but these can often be overcome.

are relatively constant, and only inspections practices are known to be different (either across jurisdictions, or because they have changed), then one can tentatively posit that the differences in inspections practices *may* be the explanatory factor. When comparing across jurisdictions, this means taking as much as possible countries that are similar in most relevant respects (economic profile, social and political structure, technologies, even natural conditions, etc.)[386]. When comparing across time, it means making sure that none of these major factors have changed – and, because in most cases they *have* indeed changed (in particular technology, economic structure etc.), comparisons across jurisdictions are more frequently possible. An alternative, sometimes interesting approach is to take jurisdictions that are significantly different (thus making the comparison clearly imperfect), but have extremely contrasted inspections practices, and outcomes that are also different but in the opposite direction from what certain models would predict. This may yield important lessons about the need to refine or qualify such models.

There is, of course, a major limitation as to what such comparisons can yield. Even if we were to have series of data long or large enough to calculate correlations (which is not the case, and we will thus not attempt such calculations), correlation is *not* causation. We cannot be claiming to prove, in any way, causation. Rather, what we are attempting to do is to find whether there is additional evidence that either aligns with what certain models and studies propose (and thus could strengthen their findings), or on the contrary lead to question or nuance some of them. What we hope for, is that the accumulated evidence may, without yielding certainties in any way, at least suggest fruitful directions for both research and practices.

ii. Compliance promotion and discretion – legal questions

Instrumental and expressive visions of the law – can tensions be resolved?

These considerations on compliance promotion and the relative effectiveness of different approaches were made from a strictly utilitarian, instrumental perspective. Such an approach is well summarized by Hodges (2015), who writes: "the purpose of regulation is to affect behaviour and performance. The purpose of 'enforcement' should be to address issues of behaviour and performance, not simply to impose sanctions in the expectation that they will affect behaviour" (p. 26). Hodges himself acknowledges that there are other principles and issues at play when considering enforcement (and tort law). Enforcement should "censure" certain actions: "it remains important for an ethical society, which supports people having respect for the prevailing moral norms, that certain behaviour should be declared to be socially unacceptable and to 'deserve' the imposition of criminal sanctions by the state (alone)218 as retributive censure for a wrongful act, and that some sanctions should be proportionate to the seriousness of the unacceptable acts" (p. 26). And tort can have a role for "securing compensation" (p. 3) – even though Hodges concludes it is highly inefficient at this task and should be generally replaced by administrative compensation systems (p. 7).

The difficulty is first, of course, that it is not that easy to define which are the cases which are serious enough to "deserve" criminal sanctions – but there are approaches to this aim, combining risk assessment and intent of the actions, and we will discuss them in the third section. The challenge to such a viewpoint is more fundamental, and comes from those putting forth a fully different vision of the law, one which is anchored in different *values.* Hawkins (2002) refers to such approaches as reflecting an "expressive" conception of laws. Ashworth (2000) is one of their exponents, and defines it thus: "my conception of the criminal law gives

[386] An example that we use in the third section is comparing Britain and Germany. Evidently, there are major differences between the two, but there also are major differences *within them*, between different regions and localities, which in some cases may well be greater than the differences in averages between the two countries. Economic structure, social patterns, etc. are all indeed different between the two – but, as we will argue further in more detail, considerably close when comparing them to most of the rest of the world.

primary place to its censuring function (…) which should be exercised in as fair and non-discriminatory a manner as possible". Scholars which consider the *expressive* value of laws as fundamental tend to take exception to the "responsive regulation" approach (and later approaches building on its fundamental vision of a need to *differentiate* the regulatory response). Yeung (2004), for instance, writes that: "the Ayres and Braithwaite model (…) overlooks the constitutional values of proportionality and consistency, which are themselves rooted in the right to fair and equal treatment". What she identifies as the key tension between her perspective and the responsive regulation approach is that the latter adopts as "reference point the goal of effective future compliance, rather than the nature and seriousness of the defendant's violation". By contrast, Yeung (and others) consider that individual rights (to a fair treatment etc.) should take precedence over effectiveness considerations. She suggests that, in fact, the responsive regulation enforcement pyramid may conflict with "the requirements of procedural fairness" (which in addition would mean that, even from a utilitarian and instrumental perspective, the approach would have problems since it could weaken one of the compliance drivers).

How much should we be concerned about such values-based concerns? We would argue here that some of the tensions can be decreased by looking more closely at how a compliance-focused enforcement approach works – but that not *all* tensions can be removed, as some fundamental divergences will remain. Looking more closely at Ashworth's arguments, he in fact makes the case for his approach *in the realm of the criminal law* – not for all types of regulations. While in a specific context such as the UK's many regulatory offences are indeed "criminalized" by statutes (but in fact rarely prosecuted, cf. Hawkins 2002 *et al.*), in most other countries the majority of regulatory offences is covered by lesser *administrative penalties* (reflecting differences between common law and civil law countries, to a large extent – even though administrative sanctions are being gradually introduced in the UK as well, cf. Tilyndite 2012 *et al.*). Ashworth states that he does not "suggest that the prevention of harm is irrelevant to criminal law: it remains significant as a fundamental justification for having a criminal law with sanctions attached". He further suggests that the problem may be the over-reliance on criminal law, whereas there are "a range of initiatives in social, criminal and environmental policy" that could be used for the "prevention of harm". His recommendation is that "the aim should be to produce a set of criminal laws that penalise substantial wrongdoing and only substantial wrongdoing, enforcing those fairly and dealing with them proportionately".

These statements by Ashworth are not necessarily in contradiction with Voermans's assertion that "rules and regulations that are not systematically observed are – in the end – pointless and futile. The overarching aim of all regulation is to have an effect on (social, economic, or institutional) behaviour" (2014, p. 42). The main difference may be that legal scholars who consider consistency and proportionality to be too fundamental to suffer "modulations" as part of a responsive approach would contend that inadequate criminal laws should be repealed, rather than enforced in a "flexible" manner. Proponents of an instrumental approach, by contrast, may contend that perfectly designed laws and regulations will never exist (even assuming that best efforts are made to improve them, the impossibility to achieve an "optimal precision" of rules has, as we have seen, been rather convincingly demonstrated). In such a universe of imperfect rules, where discretion is unavoidable, we should seek to *structure* discretion in a way that is as *effective* as possible. Effectiveness is indeed doubly important: first because *effects* are precisely what the rules are adopted to achieve, and second because if they are ineffective "the authority of the legal rules themselves may be compromised" (Voermans, *ibid.*). There is thus a values-based, rule of law case to be made *for* inspections and enforcement approaches that target improved compliance – because ineffective laws undermine the very idea of the rule of law[387].

[387] And, we would add, there are many examples of criminal legislation which have consistently and fully failed at their stated goals, and indeed produced major side effects that go counter to these goals, such as drugs criminalization – and remain nonetheless on the

Looking more closely, the is partly one of whether, and how, to apply discretion. If we accept that there is no "optimal precision of rules", then discretion is unavoidable as rules will always require some interpretation, except if they are so narrow as to become essentially useless, and even counter-productive in many instances (cf. Diver 1983, Baldwin 1995). Even if one were to attempt and remove as much discretion from inspectors' and other officials' hands, judicial discretion would remain in considering cases. The question then becomes how to structure this discretion, how to "frame" it. As Bardach and Kagan (1982) put it, "while there are powerful (…) reasons for regulators to treat all regulated entities more or less "alike" (…) under certain conditions it may be possible to justify *dissimilar* regulatory treatment" to achieve "more reasonable regulation" (p. 92).

There are, of course, scholars (and policymakers, judges etc.) who would contend that decisions can be made essentially "without interpretation" of the rules' meaning, at least in most cases. A particularly famous proponent of this view is Justice Scalia of the US Supreme Court, who wrote in *The Rule of Law as a Law of Rules* that it is essential to avoid "uncertainty regarding what the law may mean" (Scalia 1989, p. 1179). While cautious to admit that there will always be cases where "legal determinations that do not reflect a general rule" cannot be avoided (pp. 1186-87), he nonetheless advocates for making decisions as much as possible that "adhere closely to the plain meaning of a text" (p. 1184). In spite of reading very pleasantly, and of some of its caution, we think Scalia's "originalist" or "textualist" vision does not stand close scrutiny. Following Black (1997), there is good reason to think that rules are essentially "indeterminate" because of the limitations and nature of language, even before considering the questions of their anticipating situations while being unable to predict all future events, and of the social context through which their misleadingly "obvious" text has to be understood. Other scholars, commenting specifically on Scalia's thesis, have shown its weaknesses. Strauss (2008) for instance writes that "he choice between rules and discretionary standards confronts legislators and regulators routinely. It also confronts judges, or at least Supreme Court justices. The *Rule of Law as a Law of Rules* is an elegant and appropriately cautious defence of the position that rules are, as a general matter, superior" but adds that "rules in constitutional law, like many other things in the world, are most often the product—the ongoing, unfinished product— of evolution" (p. 1013) – meaning that they cannot be derived from the "plain meaning" of the legal text. Solum (2002) takes a more radically critical view[388] and writes: "The rule of law does not require a law of rules; nor does a law of rules guarantee the rule of law. The problem of judicial constraint is not that simple, and the strategies that are adequate to advance the predictability and uniformity of the law defy easy summary. The rule of law requires sound practical judgment by judges of integrity" (p. 23).

Let us conclude this short discussion by acknowledging that tensions between conflicting views of the law, and of its enforcement, can certainly not all be reconciled. There will remain a side of the debate where the preference is for consistency and predictability, and which holds the discretion can be minimized, if not abolished. This does not mean that the instrumentalist vision of regulation advocates unbridled discretion, quite the contrary – but that it holds discretion for unavoidable, and thus considers that it is best addressed by embracing it, and trying to give it a transparent and predictable framework (to the extent possible). We would also argue that such a framework should also try and give some guidance on how to determine the facts themselves, for facts are often no more "obvious" than the meaning of legal texts[389].

---

books. Thus, laws that are designed with a purely expressive approach, and without consideration for an instrumental perspective, tend to be deeply problematic.

[388] But definitely not because the author would be instrumentalist – he in fact writes about the "vice of instrumentalism" in judicial decisions (p. 23).

[389] The determination of facts is a problem more frequently addressed in a judicial perspective, but is in fact often a serious issue in regulatory matters. From our original training as a historian, we have learned that "facts" in human matters are highly problematic –

Framing discretion

Administrative discretion – theoretical overview

In spite of the many real differences between different legal traditions and systems, we would argue here that *to a first approximation* all major jurisprudences allow for a degree of administrative discretion, and it is more a matter of how it is defined and bound, than of *whether* it exists. As Bardach and Kagan (1982) showed, even when an administrative agency purports to be enforcing very detailed rules strictly "by the book", it is often impossible in practice, and "the needed flexibility, in such agencies, traditionally is attained by not enforcing the rules literally" (p. 37)[390]. Certainly, there are cases where officials and judges *refer* to fundamental legal norms that are country- or system-specific to justify or explain the refusal or reluctance to use certain forms of discretion (cf. Rothstein, Borraz and Huber 2013 for the specific example of France and Germany in relation to risk-based discretion). Similarly, we have heard from government lawyers in countries as distant as Mongolia and Ukraine that it was "impossible" to give discretion to inspectors to enforce certain norms and not others. The question of what discretion covers, and of what is allowed within it, is thus a relevant one. Unbound discretion leads to serious problems of lack of consistency (cf. Bardach and Kagan 1982, pp. 86-87), which are to be balanced against the benefits of flexibility. While we certainly cannot treat the issue here in its full depth and complexity, we will nonetheless attempt to set down a few markers.

As Voermans (2014) writes, "the duty to implement and enforce laws is generally perceived as something required by the rule of law" but "although public authorities in the rule of law-based jurisdictions are under the obligation to implement law, and enforce it if necessary, they do not have total discretion in doing so. Implementation and enforcement activities generally need to have a basis in law as well, and the law itself sets conditions for implementation" (p. 46). In other words, the first limitation on discretion is one on *how much* the state authorities can do, *how much* power they can wield. This includes fundamental principles such as *nulla poena sine lege priori* (non-retroactivity of laws), *lex certa* (clear definition of what is prohibited) and proportionality (*ibid.*). While there are many countries where these principles (while they may exist on the books) are not respected in practice, there is no disagreement among scholars as to the *legitimacy and appropriateness* of these limits on discretion – what we could call "ceilings" on what state officials can do. Rather, what is cause for disagreement are the limits on *how little* the authorities may do without violating their duties, of what would be the "floor" on discretion.

If one looks at practices, it is clear that there is essentially no case of absolute, full enforcement of any law – simply because means for enforcement are inherently limited. Even in the case of murder, for which Tyler (2013) points out that this is where deterrence can (at least in many countries) work most effectively (in principle) because elucidation rates are high (because society has agreed to allocate massive resources for each case), elucidations are clearly not 100%, and police resources *are* limited. This is even truer for other violent crime, and considerably more true for non-violent crime, and many regulatory issues. Thus, *de facto* the state exercises "downwards discretion" in not inspecting and enforcing "everything, all of the time" – because it would be impossible. One can also frequently observe that governments delay preparation and adoption of secondary legislation, in countries where it is absolutely needed for laws to function, in many

---

the naïve confidence of the "positivist" school having long been set aside (see e.g. Delfau 1978 on the evolution of historical thought, through positivism and away from it).

[390] Quoting a 1972 article by P. Schuck: "The inspector is not expected to enforce strictly every rule, *but rather to decide which rules are worth enforcing at all*. In this process, USDA offers no official guidance, for it feels obliged, like all public agencies, to maintain the myth that all rules are rigidly enforced" (in Bardach and Kagan 1982, p. 37). If such a picture has more general validity (which our experience suggests), then discretion is unavoidable, and trying to negate and repress it only makes it more arbitrary – acknowledging it openly allows, by contrast, to give it a clear, transparent, consistent foundation, for instance risk proportionality.

cases presumably because of lack of resources (both for drafting and for enforcement)[391]. If, in practice, "less-than-complete" implementation of laws is commonplace, remains the question of whether it is *legitimate.* This, in turn, can be examined both through legal doctrine, and through the possibility (or lack thereof) to *sue* the government (or any agency reporting to the executive) for inadequate enforcement of a statute.

At its root, the question is a constitutional one – and the very existence of an *executive branch* supposes that it has some power that is *independent* or at least *distinct* from the legislative one. In this sense, discretion is consubstantial to the existence of an executive power. Moving to specifics, however, the question appears far less clear-cut. It would reach far beyond the scope of this research to consider it seriously in a number of different constitutional and legal settings. What we will attempt to do is just give brief illustrations of why there is reasonable basis to consider that, in most contexts, there will be sufficient room for discretion in existing legal norms and principles to accommodate the enforcement practices that pertain to "risk-based inspections" and "smart inspections and enforcement". Investigating in more depth to what extent, and through which legal means, this can effectively be done in each given jurisdiction will be a task for further research.

In France, the question of administrative discretion corresponds to the "*pouvoir d'appréciation*" – which is foreseen by some laws, and not by others (or can be made necessary because several different principles are in conflict[392]). Administrative courts have the power to review administrative decisions (including, possibly, decisions "not to act") – and the administrative jurisprudence of the *Conseil d'Etat* has established principles that define and limit (in some cases) the ways in which the executive branch and administrative bodies can exercise discretion. When the applicable law or other norm has vested the public administration with a "bound competence" (*"compétence liée*") then there is no discretion – and the administrative courts will invalidate any administrative decision that did not strictly implement what the norm required. By contrast, when applicable law gives "discretionary power" (*"pouvoir discrétionnaire*"), the control by administrative courts will be more limited[393]. While in earlier times judges used to refuse to exercise strict review for decisions pertaining to an area of discretionary power, case law has moved towards a control of whether the public administration did not commit a "manifest error of judgement" ("*erreur manifeste d'appréciation*"), in other words a control that is not only of legality, but of opportunity[394]. Typically, administrative judges will defer to administrative decisions in cases that are highly technical. In some cases, judges apply a strict scrutiny, looking at whether the decision taken is overall proportional to the *costs and benefits* of the situation. In such situations (which, overall, are quite rare), judges in practice replace the administration's discretion with their own[395]. In some instances, the *Conseil d'Etat* has done so in order to substitute a *stricter* or *harsher* decision to the administration's relatively more flexible one[396]. From this short summary we can conclude that: (a) in a number of cases, administrative discretion indeed is present (basically, every time it is not excluded by the wording of the law) – (b) administrative case law takes into account cost-benefit and proportionality

[391] This is relevant e.g. in France, where many laws simply cannot be enforced without the additional level of precision given by Cabinet decrees (and this duality is foreseen by the Constitution). Since most laws adopted by Parliament are the reflect of a strongly executive-led majority, the frequently observed delays are not typically the reflection of political splits between Cabinet and Parliament, but of sheer overload (driven also by excessive legislative "production").

[392] There is for instance a directly applicable constitutional principle of "reconciling the protection and valorization of the environment, economic development and social progress" (Tifine 2014, Second Part, Chapter 1, Section I – accessed on 30/8/2015 at http://www.revuegeneraledudroit.eu/blog/2013/08/21/droit-administratif-francais-deuxieme-partie-chapitre-1-section-i/#.VeLnYtLS2zk)

[393] Cf. Tifine 2014, Second Part, Chapter 2, Section II, par. I – accessed on 30/8/2015 at http://www.revuegeneraledudroit.eu/blog/2013/08/17/droit-administratif-francais-deuxieme-partie-chapitre-2/#.VeLn4tLS2zk. In this latter case, in fact, a first type of error ("erreur de droit") would be for the public administration to disregard the fact that it had, in fact, discretion.

[394] Cf. Tifine 2014, Second Part, Chapter 2, Section II, Par. II A

[395] Cf. Tifine 2014, Second Part, Chapter 2, Section II, Par. II B

[396] See e.g. Tifine 2014, *ibid.*, sub-point (b) – and Eliakim (2013) (chapter *Pour quelques centimètres de trop*)

considerations, at least in a number of cases and (c) in some instances, administrative judges will overrule the public administration with their *own* discretion. Moreover, if we consider not administrative but *criminal* law instead (in the cases when regulations foresee criminal liability for some violations, which is decidedly less common in France than in the UK, for instance), the discretion *not to prosecute* is even clearer, and is a fundamental principle ("*principe d'opportunité des poursuites*"). When the public prosecutor is informed of facts that "constitute a violation", the prosecutor decides "whether it is opportune" to either "initiate prosecution" or "initiate an alternative procedure", or "to close the case" (Art. 40-1 of the Code of Penal Procedure)[397]. While there can be an appeal of this decision (to a higher ranking prosecutor), and while civil action is not excluded, this prosecutorial discretion is not limited – not initiating a prosecution is purely a matter of judgement.

The oft-stated difference between "Continental" or "Civil Law" systems and "Anglo-Saxon" or "Common Law" systems, which in any event is generally a woeful over-simplification[398], appears of little relevance to the matter of discretion, at least as far as the *principles* are concerned (the mechanisms of action and litigation being, evidently, country-specific). First, administrative discretion is grounded in the principles of *comity* and *deference.* The first "is the respect that a public authority ought to show for the work of another public authority", and is in a way nothing else than "respect for the separation of powers" (Endicott 2015, p. 20). The second derives from comity will posit that "it takes some special reason for the court to interfere with [a given] decision maker's answer to" the initial question at hand (*ibid.*, p. 234). Deference requires to pay attention to the "legal allocation of power", "expertise", "political responsibility" and "processes", four reasons for which the initial decision-maker may be in a better position to decide than the court (*ibid.,* p. 234-235)[399]. In spite of this, however, there *are* situations when the "presumption of non-interference by courts" (*ibid.*) can be over-ruled. This involves situation which are not defined as *discretion* but as *arbitrary*, *abuse of power* or *unlawful exercise of power*. Different criteria can be applied, which include: "fraud and corruption", "bad faith or malice", "use of a power for a purpose that is contrary to the statute" and taking into account considerations that are "irrelevant" to the purpose of the statute being enforced (*ibid.,* p. 230). An alternative "check-list" includes: "error of law", "irrelevance" (of matters considered in the decision), "absurd" decisions and "bad faith" (*ibid.*, p. 239). The criteria are not unlike those used in France, including the "absurd decision" criterion which is similar to the "*erreur manifeste d'appréciation"*: "if the judges are able to say that no one *in the position of the public authority* could present the action in good faith as a genuine exercise of their discretion, then the judges can interfere (…) with no breach of comity" (*ibid.,* p. 237). Deference applies to judicial review of decisions that are political in essence (e.g. budget decisions), which are either excluded ("non-justiciable", e.g. an Act of Parliament[400]) or deserve "massive deference" – but it also applies to "administrative" decisions in the narrower sense, as long as the authority making them is vested with some discretion. The latter can arise from a number of situations: "express discretion" and "implied discretion" arise from the wording of a law that gives specific powers to a decision maker (either *expressly* giving discretion, or leaving the power to act or not open, i.e. giving it *implicitly*), while "inherent discretion" relates to a power that "is essential if the body is to carry out its role" and "resultant discretion" arises when the wording of a statute is sufficiently vague to require a substantial degree of interpretation (*ibid.,* pp. 243-245). From a regulatory perspective, it is worth noting that courts "defer massively" to administrative authorities e.g. in matters of planning (*ibid.*, p.

[397] Accessed on 30/8/2015 at http://legifrance.gouv.fr/affichCodeArticle.do?cidTexte=LEGITEXT000006071154&idArticle=LEGIARTI000006574935&dateTexte=20150830

[398] Of course, one should note that French Administrative Law is in fact very similar to Common Law in its approach: it is nearly entirely based on Case Law, and relies on sets of fundamental principles, rather than on written law.

[399] For an illustration of practical decision-making by courts on this basis, cf. Endicott 2015 p. 233.

[400] See on non-justiciability Endicott 2015 pp. 251-260

262) as well as towards prosecutors when it comes to the decision to prosecute or not (*ibid.,* p. 266[401]). Hence, "typical" regulatory decisions (planning decisions by local inspectors, decisions to prosecute by HSE inspectors) are mostly covered by strong deference to the officials' discretion.

A concrete example of "discretion in dispute"

We have seen that, in countries apparently as different as the UK and France, there are in fact quite similar principles at play when it comes to administrative discretion and possibilities of judicial review thereof. Deference is the norm, but there are exceptions to it, and principles for screening and reviewing are relatively close. This leaves us with apparently quite a solid basis for discretion, including the discretion *not to act*. Let us consider a final example, a more contentious one, to see if it can strengthen our findings. Recently[402], President Obama decided to in a way sidestep Congress on immigration policy, due to the impossibility to forge a bipartisan compromise, and to act in this matter entirely on the basis of executive discretion. Not, however, the *individual,* prosecutorial "bottom-up" discretion of officials in charge of making case decisions, but *structured,* "top-down" discretion, through instructions from the Secretary of the Department of Homeland Security). The President (through his Secretary) did not, of course, instruct immigration agents to stop all actions against illegal immigrants – but he established "expansions of deferred action, with guidelines for when someone should be protected; and the new "clear guidance" enforcement memo, which lays out much clearer, and more restrictive, guidelines for when someone should be deported"[403]. This elicited of course sharp reactions from Republicans, and scholars on the conservative side of the ideological divide. This is, in several ways, an "extreme" case: first, because it corresponds to a very hotly debated issue, rife with ideology and electoral interests, and thus one where it can be expected that scholars on all sides will "push" their argument as far as they can. Second, because it relates to a "discretion-framing policy" that is particularly sweeping in its scope and strict in its guidance – and particularly in its guidance *not to act*. .

Considering this, it is striking that, if we look at the arguments made *against* the policy, they are quite moderate and limited *in substance* (if not in tone). The *Heritage Foundation*'s John Malcolm (2014) thus starts by writing that the President has a "constitutional duty to enforce the law" that derives from the Constitution's stating that "the laws be faithfully executed" (Art. II, sec. 3) – and that the Supreme Court "Court determined that the President must carry out all of the objectives and the full scope of programs for which budget authority is provided by Congress" (p. 2). He fully acknowledges prosecutorial discretion, but argues that this, "with respect to an executive's enforcement duties is based on equitable considerations in an individual case or a small set of cases" – and "is designed to help achieve statutory objectives— which in this case would include promoting the integrity of the U.S. legal immigration system and deterring violations of our immigration laws—not to frustrate statutory objectives or to effectuate a change in policy" (p. 3). Thus, he argues, since prosecutorial discretion should be the exception (and aligned with the aim of the statute), the announcement that it will be used in a sweeping, systematic way (and in a manner that the authors sees as contradicting the statute's finality) contradicts the law, and the Constitution. He goes on to acknowledge, however, the following: "this rationale may end up squeaking by in a court of law, assuming it is challenged by a plaintiff who is able to establish the legal requirements of standing" (p. 4). He adds in note the following explanation: "the Supreme Court held that the presumption against the reviewability of discretionary enforcement decisions can be overcome "where the substantive statute has provided guidelines for the agency to follow in exercising its enforcement powers" and that an agency might be subjected to a more

[401] Endicott does point out that there were some variations in jurisprudence on the question of reviewing decisions to prosecute (or not), but the latest, prevailing jurisprudence is basically full deference.

[402] Starting from November 2014

[403] Dara Lind, *The government can't enforce every law. Who gets to decide which ones it does?* Online article accessible at: http://www.vox.com/2015/3/31/8306311/prosecutorial-discretion

exacting standard of review if it "consciously and expressly adopted a general policy that is so extreme as to amount to an abdication of its statutory responsibilities." Nonetheless, no court has ever invalidated as a violation of the Take Care Clause a non-enforcement policy premised on prosecutorial discretion" (*ibid.*). In other words, while *in theory* the Supreme Court considers that discretionary enforcement decisions *could be* reviewed, it would only be possible in an exceptional case (and would require that the plaintiff demonstrates standing, i.e. that they are being harmed by the discretionary action, which may not be easy). In short, even in the United States (where the Constitution as well as jurisprudence tend to limit the executive's discretion in internal affairs), it appears that policies "framing" regulatory discretion would pass legal and constitutional muster, except in the most extreme of cases.

Even Malcolm's critiques, however little consequence they have in the end considering the case law he himself quotes, are not undisputed. Shoba Sivaprasad Wadhia, a scholar of immigration law and prosecutorial discretion issues, advances strong arguments against all of Malcolm's views. She first describes in far more details the actual *contents* of the policy in debate, and notes that in fact the guidance specifically grapples with "the more complicated cases" and thus "permits the agency to go beyond a "one-size-fits-all" approach when applying its policy on prosecutorial discretion" (p. 106). She then emphasizes the economic impossibility of full enforcement: "the government has resources to deport approximately 400,000 individuals annually—less than four percent of the deportable population" (p. 107) – which means that *in fact* full enforcement of the law is impossible, and that the practical choice is only between *structured (and consistent)* discretion and between *individual (and inconsistent)* discretion. She also demonstrates that the humanitarian basis for the new discretion policy has a long history, and that "One of the earliest documents used by the immigration agency (then called Immigration and Naturalization Service) was an Operations Instruction that allowed for "deferred action" (then called "non-priority status") for noncitizens who could show one or more of the following factors: advanced or tender age; presence in the United States for many years; need for treatment in the United States for a physical or mental condition; and adverse effect on family members in the United States as a result of deportation" (p. 109) – i.e. criteria very close to today's. Finally, she gives a very different summary of the case law, quoting the Supreme Court's earlier recognition that ""[a] principal feature of the removal system is the broad discretion exercised by immigration officials. (…) Federal officials, as an initial matter, must decide whether it makes sense to pursue removal at all" (p. 112). She adds that the "Take Care Clause" in fact has been repeatedly understood by the Supreme Court to include "broad discretion" in enforcement (*ibid.*) – and that in fact the Immigration and Nationality Act specifically "prohibits judicial review for three specific prosecutorial discretion decisions (commencement of proceedings, adjudication of cases, and execution of removal orders), only reaffirming the delegation of prosecutorial discretion powers to DHS" (p. 113). It thus appears that, even in this most hotly contested field that is immigration law, prosecutorial (administrative) discretion is as essential as it is, essentially, enshrined in Constitutional law (and case law in particular).

The legitimacy of discretion

We have clearly not *proven* (if such a thing is even possible) that discretion in regulatory decisions, and particularly the discretion *not to act*, is possible and legitimate always and everywhere. There are evidently exceptions, limits, and ways in which this discretion is organized – and this will vary from one country to the next, with significant divergences between different legal traditions. What we think can be said with some confidence, however, is that regulatory enforcement policies (adopted by the executive branch) that organize how discretion will be exercised, including in providing guidance to individual officers on what violations can be "treated lightly", are certainly not shocking innovations, or *generally* contrary to sound constitutional and legal principles (though they can be problematic in *certain* constitutional systems).

Accepting discretion as a necessary element of risk-based inspections and enforcement also does not need to mean that accountability is reduced. As things stand, in most countries, the majority of inspections and enforcement structures are usually accountable to the executive, and only through the executive to the legislative branch[404]. This is not changed by risk-based approaches, but rather they introduce an element of clarification regarding which criteria will be used to exert accountability. By publicizing a clear methodology to guide inspections focus, and to take enforcement decisions based on risk (such as the UK HSE's *Enforcement Management Model*), an inspections agency clearly defines within which parameters it will exercise discretion, and how, rather than having the default situation of "atomized" discretion at the individual inspector level (which can, in practice, never be ruled out – an inspector can always decide s/he has *not seen* something, even if rules say it should be subject to a penalty every time, for instance). Moreoever, if inspection agencies define their goals and objectives in terms of public goods to be increased and/or risks to be reduced, they allow for far more meaningful accountability, since the executive (and, in turn, the legislative) can scrutinize whether the methods used have indeed allowed for maintained or improved outcomes, or not.

It will be a task for future research to investigate how such policies can be designed, adopted and implemented in different jurisdictions – but we believe to have established sufficiently that they are *possible and legitimate.*

In addition, there is sufficient evidence that minimizing discretion results in situations where efforts end up diverted to low-priority tasks, and/or in "minimal compliance" (Bardach and Kagan 1982, pp. 102-109). As they demonstrate, "going by the book" and treating every regulatory violation, no matter how small or inconsequential, exactly with the same attention, produces results that are not only "sub-optimal", but can be downright negative, and undermine the very objectives of regulation. As they conclude, "such diversion leads managers and compliance specialists to denigrate the inspectors, to characterize them as ignorant and legalistic nitpickers, and to resist rather than cooperate with them" (p. 104). On this basis, we can now turn to consider the contents and practice of risk-based approaches that aim at making such discretion better framed – more consistent, more transparent, and more effective. Indeed, this last point is important – unbound, unmanaged discretion also has its pitfalls. As Bardach and Kagan show, the reliance on "traditional legal structure" and prosecutorial (and judicial) discretion largely resulted, in the years before the 1970s "tightening" of regulations and enforcement in the US, in a situation of "underenforcement" (p. 40). While discretion is important to "distinguish between serious and nonserious violations, between the basically well-intentioned regulated enterprise (…) and the recalcitrant firm" (p. 39), there is also a serious downside risk of capture or simply excessively lenient approach (pp. 39-42). A well designed risk-based approach, we will argue, can offer a framework that allows the positive sides of discretion to operate, while avoiding or limiting its downsides.

Finally, it is important to point out that the degree to which executive discretion (e.g. prosecutorial discretion, but also by extension prosecutorial discretion) is generally construed as legitimate depends on the legal tradition. Whereas both in the British and American tradition, and in the French and Roman one, there is deference to the opportunity principle (the executive and prosecutors may elect *not to prosecute* or otherwise enforce if it would not be opportune, i.e. would not support overall goals of public welfare etc.), the German legal tradition (and that of all countries that build on it) does not include this principle. There, by contrast, the principle of legality (*Legalitätsgrundsatz*) would suppose that every violation is equally prosecuted. While this does not really happen *in practice*, and thus the difference between legal traditions is not that stark in fact as it is in theory, it remains that the legitimacy of regulatory discretion will not be as easily established in countries where the legality principle is the norm as in others which embrace the opportunity principle.

[404] Financial sector regulators, or other high-profile "independent regulators", can be exceptions to this rule, but they are not the focus of this research.

c. Conclusion

It is clearly difficult to conclude on a topic which presents such conflicting views and apparently contradictory findings. The large number of very valuable studies also makes it hard to do justice to the field. We will nonetheless attempt to do so, in order to provide an adequate basis for the consideration of evidence from the practice. First, we will return on the need to accept the complexity of compliance, and to look beyond simple models. Second, we will return to the "big picture", and the consideration of *outcomes*. Finally, we will see that a "modest" vision of complementary, complex compliance factors is sufficient as a foundation for risk-based inspections, and that risk can in fact be a tool that allows to move beyond some of the apparent contradictions and challenges.

i. A second look at "deterrence" studies

Let us first look back at a couple of studies specifically considering "deterrence-based" compliance. Faure and Garoupa (2005) consider the limitations on deterrence in cases where fines may fail to be commensurate to the illicit gain for a variety of reasons, and where forfeiture (of illicit gain, or of wealth deemed to come from an "illegal source", etc.) is introduced as a complement. Importantly, the authors underline that such "measures" also respond to the idea that "crime should not pay", and not only to a deterrence logic (p. 280). They also see forfeiture of illegal gain as substituting itself to compensation payments in the case of "victimless crimes" (*ibid.*), and put in the perspective of "corrective justice" (pp. 289-290). They consider the legal frameworks for such practices, including the use of *civil forfeiture* in the US[405], but considerations of effectiveness are based on *models* and *assumptions* followed by *logical deductions* – without any guarantee that they correspond to practice. The authors refer to "criminal lawyers" considering the deterrence model as particularly appropriate (p. 282), and to both economics and the principle of proportionality as requiring *marginal deterrence*, for which forfeiture of illegal gain can be a useful instrument, when combined with fines (the fines can be modulated based on the seriousness of the offence, while forfeiture provides a "baseline" bringing back offenders to the *statu quo ante* – p. 288). In fact, the authors themselves acknowledge that many criminals are (evidently) not being deterred (p. 283), but they do not really question the model. While many of their arguments of *principle* are convincing (e.g. regarding proportionality, ensuring crime does not pay etc.), these are values-based arguments. The *effectiveness* case for the deterrence side is unproven.

In a 2015 paper, Bentata and Faure consider the evidence on the activity of Environmental NGOs (ENGOs) in France, through environmental cases litigation brought before the *Cour de Cassation*. They suggest that, in a context of limited inspections and enforcement resources (p. 5), ENGOs take up important cases (in terms of environmental damage) that would otherwise be left out – because the regulator is focusing on the higher-risk, larger-size entities, and individual damage is too small to lead to private litigation. They further show (pp. 6-7) that ENGOs focus on cases with a high impact on the environment rather than on "personal nuisances"

[405] The civil forfeiture practice in the US has come under increasing criticism in recent times for the manyfold abuses it has led to, with weak rules of evidence and perverse incentives leading to police departments routinely abusing their powers. There is a growing, and increasingly bipartisan consensus that the practice should be ended – see e.g. concurring conclusions from the American Civil Liberties Union (https://www.aclu.org/issues/criminal-law-reform/reforming-police-practices/asset-forfeiture-abuse), the Cato Institute (http://www.cato.org/events/policing-profit-abuse-civil-asset-forfeiture), libertarian writers such as Radley Balko (https://www.washingtonpost.com/news/the-watch/wp/2014/05/19/new-media-investigations-show-that-the-asset-forfeiture-racket-is-still-humming/) etc.

issues (ENGOs focus more on water issues, private cases on noise and soil, relatively speaking). ENGOs and enforcement through court cases thus appear as meaningful complement to state regulatory inspections (p. 11), but again the question of *effectiveness* is not fully investigated. The authors show that, over time, defendants' overall compliance rate with safety measures (as evidenced by court proceedings) has increased, following an increase in ENGO litigation, but this is at best correlation between two trends – and the fact that there is litigation suggests that this compliance was insufficient to ensure environmental protection. Thus, the ENGOs' role appears potentially meaningful, but within a broader concept of deterrence which remains unproven, at least within the paper.

Rousseau, in a 2007, considered closely a dataset on environmental inspections and enforcement in Flanders, which gives the possibility to look at correlations between compliance and inspections/enforcement more closely, particularly given the relatively long time-series, and the repeated inspections of each establishment[406]. The model used to investigate enforcement effects is squarely rooted in the "deterrence" vision (p. 2)[407]. Rousseau summarizes her findings as confirming the deterrence effect of increased inspections, but not that of sanctions. She discusses the fact that the agency uses sanctions relatively rarely, and that the level of fines remains far lower than what the legislation authorizes, i.e. the fully "enforcement pyramid" is not really being used (but the *threat* of the pyramid's "top" is likely to be used – cf. pp. 8-11). On the *interpretation* of results, we feel like there are important points that could be seen differently from the author. First, she outlines factors increasing likelihood of inspection on p. 17 which, in fact, squarely show that the agency is using a risk-based targeting approach - meaning that, if targeting is done well, one would precisely expect (i) a relatively high percentage of violations (which indeed is found) and (ii) some effect of inspection and enforcement visits (which, again, seems to take place). Thus, the findings may not really reflect the effect of inspections *overall* but of a *targeted risk-based project*. Second, and most importantly, the fact that the increased inspections programme seems to have an effect, but sanctions do not seem to have one, may suggest that the effect is *not* (or not entirely) linked to *deterrence*. It could very well be that the repeated, extended (longer duration of visits) interactions have allowed to increase the inspected businesses' knowledge, and to build a trust relationship where persuasion has played a significant role. Finally, and relatedly, the fact that inspectors and courts do not use the full scale of sanctions available, and impose (when they do) sanctions that tend to be far lower than marginal abatement costs for major violations, again challenges the "deterrence" approach – this time in a "feasibility" perspective. One can assume that both inspectors and judges are not ignorant of the problem – but imposing massive financial sanctions on businesses, while it *may* increase the general deterrence effect (which may or may not a really important driver of compliance), would surely pose serious financial hardship to the sanctioned enterprises. This could in some cases put them out of business, or at least threaten their viability, and in the meantime make it *even more difficult* for them to invest in the required pollution abatement equipment. Thus, overall, while the finding that this specific inspections project was successful at increasing compliance appears robust, the *reasons* why it was so are probably more complex than suggested.

---

[406] Two remarks are required. First, the group under consideration is a high-risk group, specifically targeted by the environmental inspectorate as part of a "project" – which translated into more than 3 inspection visits per entity and per year, on average (see p. 7), which is a very high number, and thus makes it difficult to assume that findings can be easily generalized. Second, the findings incidentally show the problems with the notion of "compliance", because so many of the non-compliances are administrative rather than substantial (*ibid*.), i.e. are non-compliances without a direct environmental impact (and, for some, without even an increased *risk* of harm).

[407] The introduction includes a short literature summary. It includes a point on Nadeau's 1997 findings that inspections and enforcement actions reduce *the length of time spent in non-compliance*, and that enforcement has a stronger effect. This is an extremely unsurprising finding, we would say, and very different from a conclusion on relations between enforcement actions and compliance *overall.* There is little doubt that, if you are inspected (maybe repeatedly) and sanctioned (again, possibly repeatedly), this is likely to push you to start putting yourself in compliance. The question is whether controls and enforcement actions are the most effective approach to increase compliance *across the board* among all regulated entities.

ii. Taking a more modest and nuanced approach

While we have collected examples suggesting that the deterrence model is generally unproven, it is not to single out so much this factor, as because it has been the model used the most uncritically across many studies. Scholars investigating psychological drivers such as Tyler *include* deterrence, while suggesting that it may be weaker than e.g. legitimacy – but many deterrence-based studies barely acknowledge (and then proceed to ignore) other drivers. This conclusion by Parker and Lehmann Nielsen (2011) seems to us highly appropriate: "the range of factors that are hypothesized to influence compliance are so complex and interrelated that it is very difficult to holistically test them all, or even to clearly hypothesize how they interact and in what direction causation flows" (p. 6). Likewise, in their summary of contributions on "Effective enforcement of consumer law in Europe", van Boom and Loos (2007) conclude to the importance of a multiplicity of complementary approaches. They challenge the idea that litigation against an infringing firm (even successful) necessarily leads to a change of behaviour (p. 6), and cover several examples of effective interventions based on information, informal pressure etc., rather than formal enforcement (p. 4). They also see merit, however, in systems which enable *group action* (with important nuances compared to US *class actions* – cf. pp. 5-10). Their concluding view is that self-regulation (with or without a state regulatory "backstop") and public supervision and enforcement are complementary and not contradictory, and that group action can be a useful supplement to both (pp. 10-11), a view that would fit well with a view of complex (and evolving) compliance factors.

In a 2007 paper, Voermans talked about the "aspirin-like effect of sanctions", suggesting that (just like for aspirin or, say, homeopathy) many people will assert that "it helps", without being able to explain why or how (and without, it goes without saying, scientific evidence thereof – cf. p. 59). He considers the problem of laws that "do what they are meant to", of rules "that are functional", as central (p. 57), but the question of what mechanisms lead from rules to behaviours as very much still unsolved. Indeed, while voluntary compliance is preferred, we know it does not always happen – conversely, while no one really doubts that enforcement has *some kind of effect* on compliance with rules, how, and how much, are other questions (p. 58). The *assumption* that more control and more enforcement will lead to better results has led to what he sees in the late 1990s and 2000s as a considerable increase in inspections and enforcement efforts, in particular on the part of local authorities – involving more professionalism, but also a number of new (previously unheard of) enforcement directions – all without much basis in evidence (p. 56-60). Rather than looking at the *logic of motives* behind compliance, these measures have followed an *administrative logic* – the more is done, the more it is expected to be effective. Voermans considers both large-scale, high-level data, and findings from psycho-social studies. Data first: quoting van Velthoven, he shows that the chances of being caught, and the potential fine, are so vanishingly small that it is impossible to plausibly explain widespread legal compliance based on deterrence (pp. 61-62) – even though, of course, in specific cases, targeted and focused deterrence may be effective on specific persons. Findings second: studies find that most people appear *not* to be motivated by the fear of sanctions (but by values), but that on the other hand they think *others* are motivated by calculation and fear (deterrence). One reading could be that we want laws and sanctions (and make them, if "we" sit in Parliament) for "others" – based on very much unproven conceptions of what drives behaviour. Another (not incompatible) reading is that compliance is complex, and that "we" may be in both positions, successively or at the same time: of complying because of values, or because of fear. Just as much as we cannot dismiss the fact that enforcement surely has *some* effect, it is clear that most compliance *cannot* be explained through deterrence. A vision of complementary compliance drivers, of varying importance according to contexts and groups affected, is the best we have.

In addition, there are good arguments to be made that compliance should not be the *only* objective of enforcement activities and mechanisms, and that there are legitimate value-expression issues that should be

considered. We have seen how this is relevant in the case of illegal gain, for instance. Yeung (2013) stresses the importance of balancing the effectiveness considerations of "better regulation" approaches with "constitutional values, including transparency, accountability, due process and participation" (p. 3). She also cautions that many of the more "efficient" or "responsive" sanctioning and enforcement approaches proposed may conflict with key principles of criminal law ("censuring the wrongdoer", sanctions entailing "serious consequences" and "moral stigma" – but also "procedural safeguards"). We have also noted above Yeung's concerns about the tensions between proportionality and responsive regulation. All these are important, and valid – reminding that effectiveness cannot be the sole consideration. In fact, from a procedural justice perspective, we would argue that not properly considering these values would in the end probably *harm* effectiveness. This further reinforces the case for a complex and balanced vision.

iii. Using "risk" to overcome (some) problems and tensions in models and theories

As a transition to the next section, we would like to point out the way "risk" can be a powerful tool to overcome some of the tensions and problems in compliance theories and compliance-promotion models. As we have indicated above, risk-based targeting is quite possibly the reason why the environmental inspection project that Rousseau (2007) studied yielded rather convincing results. More generally, modulating inspections and enforcement approaches in relation to risk is an "ideal" complement to the responsive regulation approach, as Baldwin and Blanc (2007, 2010) have already noted. We see the relevance of "risk" as coming from two perspectives: a legal one, and an effectiveness one.

On the legal side, risk can be an instrument on which to base the application of the key principle of *proportionality*, that Yeung for instance is worried can be harmed by a purely responsive approach to enforcement. In a risk-based approach, enforcement measures should always be proportional to the risk caused by the violation(s) found. The behaviour of the business operator, which is key in the responsive regulation approach, can be integrated as one of the *risk dimensions*, alongside the inherent hazardousness of the activity, and the severity of the violation. Thus, responsiveness remains, but on a foundation of risk proportionality.

On the effectiveness side, whatever the combination of compliance factors and drivers, risk-based *targeting* can be a way to optimize the intervention. It should help minimize the intrusiveness of inspections and enforcement where they are little needed (thus rating well from a procedural justice perspective, and minimizing resistance to voluntary compliance), while intensifying contacts where they are most needed – not only from a deterrence perspective, but also from a "quality of the regulatory relationship" one (more time and attention on cases which need it, meaning also more advice and time to create trust where possible). At the same time, if the balance between different risk dimensions is properly done (i.e. targeting incorporates both probability of a violation, and potential severity of its effects), risk-based targeting can maximize the effectiveness of deterrence effects (by focusing this deterrence on where it will yield most results). Before considering practical cases, and how much these optimistic expectations hold up, we will now consider the existing literature on risk and regulation and what it can bring to our understanding of risk-based inspections.

3.3. Risk and regulation – definitions, debates and issues

a. Defining and measuring risk

If, as we have suggested above, risk may be a potentially useful instrument to overcome internal tensions and contradictions in compliance strategies, and to help give discretion a sounder foundation, we first need to have as much as possible a clear understanding of how to define, and measure, risk. Evidently, considering how polysemic the word is, and how widespread its use is in common language, narrowing down its meaning is not easy. Scholars who speak about "risk" in writings focusing on regulation and enforcement sometimes do it in a way that is very much open, not to say vague, and without a clear definition. Others, who research "risk" as their core subject, investigate its different meanings and perceptions among different groups, and the effects of these differences – and, in order to do so, they precisely need to leave the definition open (is "risk" what people perceive as such). By contrast, practitioners of regulatory enforcement have been working on building a definition that commands some consensus, in order to create a foundation for their work.

Both of these aspects of risk are, of course, of interest for this research. The "open ended", multi-faceted approach allows us to understand how extremely contrasted visions of what is "risky" or "dangerous" can coexist, and how they shape the emergence of regulations and regulatory bodies. Once we move to our assigned task of trying to assess risk-based inspections practices, however, we need to have a meaningful definition of what this means, one that is not "all-encompassing". We will thus examine what it is that is called "risk" in the context of risk-based inspections. Because one of the important challenges is also how this risk should be measured and assessed, we will also briefly consider this question.

i. Risk, hazard, compliance – from "risk as likelihood of violations" to the "two dimensions" of risk

When considering research on regulatory enforcement, some of it appears to have a very narrow understanding of what "risk-based targeting" could be, equating it with targeting entities that are the most likely to commit violations. May and Winter (2012), for instance, write that "the enforcement literature is consistent in arguing that effectiveness is increased by going after the types of cases that historically have higher rates of violations" (p. 224). Though they add that there are also "other ways of identifying higher risk entities", they do not list any. Equating "risk" with "likelihood to commit violations" is of course exceedingly simplistic, and assumes that all violations are equivalent in potential consequences – or that "risk" has no other meaning than "risk of violation". Generally, looking at enforcement essentially from a "deterrence" perspective tends to lend itself to equating "risk" with "probability of non-compliance" (see e.g. Scholz 1994, pp. 426-427 for an example).

Some other studies take the opposite tack, and suggest that "risk-based approaches" consider only the potential consequences of the damage, without really looking at probabilities of violations. This is how the following remark by van der Heijden *et al.* (2015 a) could be interpreted. Looking for an explanation of why Chinese inspectors seem to target precisely those that have the strongest voluntary compliance level: "another explanation may be that agents use a risk-oriented approach to enforcement, and prioritize those farmers and types of violations that could create the largest damage. Whilst such risk-oriented approaches make theoretical sense, there is a risk of an overly technocratic implementation and too strong a reliance on the heuristics underlying these approaches" (p. 13).

By contrast Baldwin and Black (2010), who have closely studied how regulatory agencies define risk-based approaches, rightly start by clarifying that such approaches "walk on two legs": "The key components of such [risk] assessments are evaluations of the risks of noncompliance and calculations regarding the impact that the noncompliance will have on the regulatory body's ability to achieve its objectives" (p. 181). They also underline that risk-based approaches are a clear departure from regulatory visions based exclusively on compliance with rules: "the frameworks vary considerably in their complexity. All, however, have a common

starting point, which is a focus on risks not rules. Risk-based frameworks require regulators to begin by identifying *the risks they are seeking to manage, not the rules they have to enforce*" (p. 184 – emphasis ours).

The summary offered by Baldwin and Black indeed matches the practice as we have also been able to observe it in many countries. A good example and summary is offered by BRDO's 2012 *Common approach to risk assessment,* which distinguishes *hazard* from *risk.* Hazard (pp. 8-9) is the adverse effect that could arise from public welfare from given activities that are within to the regulatory body's competence – and the severity and magnitude of this hazard need to be assessed as *one dimension* of the risk. The second dimension is the "likelihood of compliance" (pp. 9-10). The *combination* of these two dimensions allows to assign a level of risk to a given activity, establishment etc. Definitions used by the World Bank Group (2013 a) and the OECD (2014) make do without the reference to compliance entirely, and rather focus wholly on the notion of "adverse event": "Risk should be understood here as the combination of the likelihood of an adverse event (hazard, harm) occurring, and of the potential magnitude of the damage caused (itself combining number of people affected, and severity of the damage for each)" (OECD 2014, p. 27).

In other words, while It is relatively uncontroversial to point out that inspecting roughly every type or size of business establishment equally is unlikely to yield optimal resource allocation (cf. Kagan 1994, pp. 409-410), it is not as easy to agree on which criteria should be used to measure risk, as this first requires to agree on a definition of risk. Our own experience working with inspectorates in former Soviet countries shows this to be one of the most difficult and essential questions – getting agreement on the fact that risk-based targeting "in general" would be better than no targeting is relatively easy, but disagreements arise when trying to define what risk-based targeting means.

ii. Several visions of risk – strengths, weaknesses and challenges

As pointed out, the notion that "risk" is the combination of the likelihood and potential magnitude of damage caused by an adverse event is not self-evident, nor is it universally accepted, even though it corresponds to what inspectorates and regulators claiming to have a "risk-based approach" generally understand under this term. There are at least three ways to conceive risk from a practical perspective, in terms of business establishments or objects of inspections:

- Probability of non-compliance with applicable regulations
- Relevance of the type of establishment to a specific "risk type" that is seen as an important priority by the government or administration
- Combination of likelihood and potential magnitude of hazards that can be caused by the specific type of establishment, be they measured through statistical work or through more "qualitative" experience and practical insights.

These three ways of defining (and thus of assessing) risk all have their own legitimacy, but are unlikely to yield similar results. They tend to be dominant in different countries or institutions, and/or to be supported by different groups, linked not only to different worldviews but to different interests. Rothstein, Borraz and Huber (2013) showed how "risk-based approaches" (in the sense of "proportionality of regulatory response to the likelihood and potential magnitude of hazard") have had difficulty to emerge in France. A different way of phrasing the same would be to say that, in France, "risk" conceptions tend to correspond to the second type: priority areas determined (based on a variety of factors) by the government and/or public administration. We will shortly discuss here some of the most salient issues pertaining to each of these visions of "risk".

Risk as "likelihood of non-compliance with regulations"

Focusing on the *risk of non-compliance with regulations* is the approach that may seem most correct from the perspective of an *expressive* use of the law, and supported by many regulators and scholars[408]. Laws are to be complied with, the executive branch (and its regulatory agencies) are there to implement these laws, and thus inspections should aim at identifying, punishing and deterring non-compliances of all kinds. "Risk" is thus nothing else than the risk of someone not complying with norms. It is worth nothing that this tends to be the prevalent understanding of "risk" in former Soviet countries, and when inspectorates there are required to adopt a "risk-based approach", and if there is no further implementation follow-up to ensure they consider *harm* rather than *violations,* this is the one they generally follow.

Such an approach, however, has practical results, if it is followed by an inspectorate. In the former Soviet examples we have observed, for instance, when developing criteria to classify establishments in different risk categories (and subsequently plan inspections prioritizing "higher risk" ones), inspectorates start by defining "high risk" as "more likely to infringe rules". This is generally done without consideration to the importance or relevance of these rules, or to the magnitude of the potential negative impact of infringements. Since non-compliance is seen as a risk *per se*, it does not matter what type of rule is infringed, or.to what degree. This results in considering smaller businesses as systematically higher risk (non-compliances, though often minor, are most frequent there, because of lower resources and expertise), and in a focus on high-volume activities such as trade, catering etc. – where, again, non-compliance tends to be frequent but usually minor in terms of effects on public welfare[409].

In theory, one could develop a more sophisticated risk-based planning approach from a "legal compliance" perspective, using the type of sanctions that can be incurred as a proxy for the seriousness of the offences. However, this would be complex to implement *seriously* (classifying all infractions recorded, analysing where the most severe are found, etc.). More importantly, one cannot assume that the legislator had a full technical understanding of the field being regulated, and insight into what activities would potentially create the highest threats. Thus, the classification would likely remain sub-optimal in terms of achieving useful social outcomes[410]. Finally, simply because there is a vast number of regulations and potential infractions, it is not unlikely that most businesses would end up being "high risk", because many (however minor) violations can be found in most establishments[411]. Since the purpose of a risk-based classification is *targeting*, this would defeat its purpose, as the "target" would be too broad.

Experience in the FSU shows that this is indeed what happens when risk criteria are developed in this spirit (and this is made even more obvious because the regulations there lack focus and are over-detailed and over-prescriptive). In Ukraine or Kazakhstan, for instance, risk criteria for inspections developed by the Standardization agency ended up classifying the vast majority of wares as "high risk", regardless of whether any injuries or deaths were ever recorded as a result of their use.

---

[408] See May and Winter 2012, Scholz 1994 for instance – an approach that puts compliance with legal norms as the key objective is congruent with the centrality of equal treatment before the law expressed e.g. by Yeung (2004, 2013).

[409] Though Baldwin and Black (2010) rightly point out that, in some cases, there can be a "huge cumulative effect of particular types of compliance failures across firms" that the harm-based vision of risk may underestimate (p. 203). In the cases we have observed, however, the disproportion between the means employed and the pettiness of problems addressed was generally striking.

[410] On the limitations of rules see e.g. Baldwin 1995, Black 1997.

[411] This is a contentious and clearly unproven assumption but there are some pointers suggesting it may be correct. Even in the UK, where efforts are clearly being made to (a) reduce the overall "regulatory burden" (whatever one may think of whether this expression is appropriate) and (b) inform businesses about rules, regulatory agencies generally target bringing most businesses to be "broadly compliant" rather than "fully compliant", an objective they consider to be as impossible as it would be relatively useless (again, considering the vast number of rules and the fact that many of them are of little significance). In former Soviet republics, we have repeatedly heard from both businesses and inspectors that, if an inspector wants to find violations, s/he will find them, considering the myriad of confusing norms. Hawkins (2002) as well as other scholars having studied in details "enforcement styles" all concur that, in general, inspectors avoid enforcing *everything* because there is *always* some norm or other that is not being complied with.

"Politically prioritized" risk

Relying on *risks as prioritized by political programmes* (or by political, elected office-holders in general) can also claim to have a legitimacy, i.e. the political one (clearly a stronger claim in democratic regimes than in authoritarian ones). In this perspective, the executive branch is legitimate to prioritize hazards that it sees as more important. This is articulated in some EU countries (e.g. by some in France) to justify having inspecting agencies directly subordinated to ministers, and receiving direct instructions from them that "interfere" with their usual planning. The justification is that ministers (owing their positions to elections) are more responsive to citizens' concerns, and that this responsiveness is essential[412].

In the former Soviet context, such "responsiveness to citizens' concerns" is not absent, even where elections are not free – since even in authoritarian regimes, keeping the majority "not overly dissatisfied" is important for stability. Ministers or presidents frequently interfere with planning by inspection agencies – sometimes for reasons that correspond to real public concerns, but often for other reasons than safety (e.g. to increase government revenue, or target businesses associated with rival politicians, etc.). We saw, for instance, the President of Tajikistan ordering different agencies to inspect all gas stations, because (supposedly[413]) of some concerns (supposedly) with fraud, and with price increases.

The problem is that very often, instead of responding to a "real" issue[414], these are sequences whereby politicians "spin" some incident reported by the media, focus on it and proclaim a "strong" regulatory response as a solution – without the problem having been analysed, and without knowing whether inspections can in any way improve it. There is neither analysis of the real risk level , nor of the response's adequacy. In this perspective, politically-driven inspections have been conducted in Tajikistan to "respond" to increases in fuel prices (gas stations inspected), in Mongolia during discussions about foreign investment in mineral extraction (mines inspected), etc. None of these, of course, made any difference to the real issue. In theory, of course, such "politically-identified" risk approach could be genuinely responsive to the "perceived risk" (as defined by Slovic *et al.*[415]) of the majority of the voters – but it appears that, in both democratic and authoritarian countries, it is more often used as a way to divert attentions from problems the government is failing to solve, and give the illusion of action. Generally, the evidence available strongly supports the case to make regulatory delivery agencies more independent from direct political supervision – and the definition of "risk" independent from political intervention.

Risk defined, and assessed, in relation to probability and degree of harm

In contrast to the first two approaches, defining risk as *the combination of the probability and the possible magnitude of adverse outcomes* is more of a "technical" (or "technocratic") view. It is based (as much as possible) on science, but in the end assessments, classifications and prioritization are done by "technical specialists" rather than scientists – and risk-based approaches have to make assumptions where there is scientific uncertainty[416]. Risk is defined as what can create harm (to life, health, the environment, etc.[417]) –

---

[412] This "responsiveness" is precisely what is seen by advocates of "risk acceptance" as a problem. What one side calls "responsiveness to citizens' concerns", the other calls "risk regulation reflex" (see next section on discussions of risk and regulation).

[413] While fraud in gas stations was certainly a concern, few believed that the inspection campaign would decrease it, but rather it was seen as a fig-leaf for more rent-seeking for inspectors and their supervisors.

[414] I.e. one that would be confirmed as really significant by examination of data.

[415] See Slovic *et al.* 1981, 2002.

[416] See next section for a discussion of the issue of uncertainty in relation to risk and regulation.

[417] "Harm" is not limited to physical issues – it can be financial/economic (loss in state revenue, market distortion, etc.).

and the risk level is proportional to how likely such harm is to occur, how severe it may be and how many people it would affect (or what would be its scope in environmental or financial terms etc.).

In this perspective, inspections should be targeted at the establishments where the combined likelihood and potential harm is greatest, which means not just greater frequency of inspections, but also "deeper" inspections, with more time spent on site, more qualified staff involved etc. In the third part of this research, we will be examining the available empirical evidence of the effectiveness (or lack thereof) of such approaches. Thus, we will for the moment set aside the first challenge to such approaches, whereby opponents of "risk-based inspections" suggest that it amounts to "regulatory surrender", and results in excessively weak enforcement[418].

If we set aside this question of effectiveness, there remain two major challenges in implementing such a system based on "actually measured/assessed risk to public welfare": a technical one (how to get relevant data and how to plan in practice) and a legal one (is it legally acceptable to thus focus and "willingly neglect" what is considered as "lower risk"). Both challenges have been raised both by scholars and in practice.

The legal principles argument against risk-focus and risk-proportionality is related to the challenges made against "responsive" approaches, and rests on the idea that risk-based approaches may break equality before the law, which is a fundamental principle (see e.g. Yeung 1984 pp. 82-83, 87). While we have discussed this argument already in the context of compliance models and discretion, it is worth restating here that, to us, this is not really a tenable position when considering actual practices rather than theoretical models, at least in the vast majority of cases. Indeed, "non-risk based" inspection approaches do not show less disparity in inspection frequencies, criteria used by inspectors, enforcement decisions etc. in most cases (see Blanc 2012 pp. 21-27 for some examples – interviews with both businesses and officials in France suggest disparity of inspectors' approaches and decisions, and disparities in targeting, are very significant issues)[419]. Thus, risk-based approaches should not, in our view, be appraised against an "ideal type" of entirely unbiased inspections, but against a reality of inconsistent and sometimes incoherent practices. Rather than introducing bias, risk-based approaches can thus be seen as introducing an *organizational principle* in practices where "equal treatment" does not exist anyway[420]. .

As for the technical implementation challenge, it has two key elements:

- What parameters should the risk classification be based upon, how to measure them, and how to then "rate" establishments according to these?
- How to turn these criteria and rating systems into a functioning planning tool, in particular how to get the relevant data on establishments and manage it?

There is a trend to base risk analysis, criteria development, ratings etc. on sophisticated "data mining" techniques, using statistical tools to determine "objectively" (though the selection of the parameters being analysed is never purely objective) the most relevant parameters and thresholds. This approach is most often proposed for tax inspections planning (see e.g. chapter by Vellutini in Khwaja, Awasthi and Loeprick 2011) – and is most applicable in their case, as tax and accounting data are suited to processing through such tools.

In practice, deploying such approaches is often simply impossible, or extremely difficult. As Baldwin and Black (2010) point out, regulators may be "dealing with low frequency events from which reliable probabilistic

---

[418] See e.g. Tombs and Whyte 2010 with precisely this title.

[419] See also Badarch and Kagan 1982 pp. 67-69, *contra* Yeung, on the many unintended and negative consequences of an excessively rigid adherence to "equal treatment", "impartiality" and "objectivity" if they are not balanced by other principles.

[420] The same remarks could be made about the problems of consistency and transparency noted by Baldwin and Black (2010) in some examples of implementation of risk-based approaches (p. 204) – while their points are perfectly valid, we would still argue that risk-based approaches should be compared to "actually existing alternatives" rather than to "perfect models".

calculations cannot easily be drawn or with conditions of uncertainty in which the risk is inherently insusceptible to probabilistic assessment" (pp. 184-185). Even when the issues regulators deal with (say, food safety) could *in principle* lend themselves to data-driven approaches (because contaminations, outbreaks etc. are frequent), the quality of data makes it frequently impossible *in practice* (because detection of contaminations depends on reporting by and testing of patients, which rarely happens, and leads to considerable under-detection and bias). In addition, even when it comes to the data that inspectorates themselves could hold, the relevant data on establishments and inspections results is either unavailable in consolidated and computerized form, or incomplete and inconsistent. This is not just the case in the poorest countries of our sample (such as Tajikistan, where no data is yet computerized, except for tax data of the largest taxpayers and the main cities), but in middle-income countries such as Ukraine or Kazakhstan (where some data is available, but incomplete, often inconsistent etc.) – and for many inspectorates in the EU, even among its richest members (data might exist but not consolidated, or may be in numerous incompatible systems, etc.). Thus, in practice, such statistical analysis as the "pure" foundation of risk-based planning is not a feasible option.

In practice, there exists a workable alternative way to develop such rating systems, far less statistically rigorous, and thus introducing more bias and discretion. The essential parameters of risk for a given "sphere" of regulation and control (e.g. "food safety" or "building safety") can be determined by a group of experts (scientists and practitioners) based on (a) the existing state of science, (b) practice and experience around the world and (c) experience in-country (even if summarized more in a "qualitative" than strictly "quantitative" way) as well as (d) available data on the issues being supervised (whatever its limitations). If done properly, in our experience, the main parameters will often be agreed upon relatively easily, be rather consistent across countries, and effectively correspond to actual risks "on the ground". For instance, in the food safety sphere, key parameters to classify establishments according to risk tend to be: (i) type of products processed, (ii) types of processes used, (iii) volumes, (iv) specifics of population served, (v) prior history and track record. This corresponds to the combination of "*inherent risks* arising from the nature of the business's activities and, in environmental regulation, its location" and "*management and control risks*, including compliance record" (Baldwin and Black 2010, p. 184)[421].

In the absence of "data mining", rating and ranking based on these parameters is subject to improvement and refinement through a "trial and error" process. The group of experts developing the rating instrument will affect scores to different parameters (corresponding to different types of processes, different sizes of establishment etc.), then define overall score thresholds for classification as (e.g.) "high", "medium" or "low" risk[422] –based on practical experience and outside examples. The thresholds' levels have to ensure that establishments with only minor risk factors end up as "low", those with several critical risk factors end up as "high" etc. It is then crucial to *test* and *adjust* these scores and thresholds: the risk criteria are tested against real-life cases of establishments. If obvious aberrations occur, the scores and/or thresholds are modified. Once the system is in use, adjustments may occur if too many, or too few, businesses end up in "high risk" and "medium risk" categories. These categories are to be used to selectively allocate limited inspection resources, so the risk classification should look like a pyramid, with more in "low", less in "medium" and even less in "high"[423].

[421] A variation of this is to consider "inherent risks" as linked to the type of activity and its size, "vulnerability factors" that can increase inherent risk (e.g. location, populations affected), and "track record".

[422] Three categories of risk being the minimum, and in our experience usually a sub-optimal number. Baldwin and Black (2010) rightly remark that the number of risk categories varies greatly. See BRDO (2012) for one example of "more than three but not too many".

[423] A key "reality check" is to compare the risk categories thus created to relevant statistics on hazards affecting the country, when possible), otherwise absurdity can ensue. E.g. in Kyrgyzstan hairdressers were classified uniformly as "high risk" due to old Soviet-time

Implementing these criteria for actual planning is often another challenge, because it requires consolidated data on establishments and software to use it. Research has shown that not only are consolidated databases with adequate information rare in developing countries and transition economies, but also in many agencies of OECD countries (cf. Blanc 2012). Some of the challenges involved in setting up such systems are:

- Collecting the information initially to create a database;
- Setting up a mechanism to update this data constantly;
- "Pooling" data across inspectorates to improve efficiency and effectiveness.

The overall take-away from the experience in designing risk-based rating and planning systems is that this is feasible if one moves away from a "statistics-based" approach and adopts a more flexible one, which incorporates generally available scientific finding, aggregate data, lessons from the practice etc.. The difficulty is then mainly in the implementation, which requires data and information management. While such approaches may appear excessively "unscientific", the point is again to question to what practices they should be compared. When put against a complete absence of targeting except by the whim or hunch of individual inspectors or managers, or very crude approaches based only on individual experience of seasoned inspectors, such imperfect risk-based approaches are considerably more evidence-based and consistent.

An "alternate account"? "controlling harms" through "projects"

At least one author has somewhat challenged the terminology of "risk" altogether, and proposed an alternate account of what "control" work is about, and how to improve it. Sparrow (2008[424]) deliberately avoids the word "risk" and prefers "the word "harm" for its freshness and for its generality, and for the fact that scholars have not so far prescribed narrow ways to interpret it. I'd like to find a way that covers the broadest set of bad things.[425]" In practice, the use of "risk" is probably far more flexible than Sparrow suggests, and for all intents and purposes his use of "harm" is not very different from what we have named above "adverse effect". In fact, many of the definitions of "risk" in a regulatory inspections context in fact use the word "harm" (risk being equivalent to the combined likelihood and potential magnitude of harm). This being clarified, let us consider what Sparrow has to say about both *harms* and *risks.*

His primary concern is that broad terms such as "risk" cover a number of different "operational challenges", and that these are insufficiently investigated. First, there are both probabilistic risks, as well as current or past problems[426] – and, in most cases, regulators have to deal with both types, but they involve different "operational challenges". In addition, there can be many *levels* at which "risks" or "problems" manifest themselves and, in Sparrow's view", the "literature seems to have gravitated to the highest levels and to the lowest levels of aggregation, with less attention (so far) paid to the messy, complex and textured layers in between[427]". This is the core of Sparrow's argument – that the actual *operational* level has been mostly forgotten. He sees risk perceptions research (e.g. Tversky and Kahneman 1974, 1979) as helping us to understand reactions, decisions and behaviours at the individual level – and "at the opposite extreme – the highest levels of aggregation – risk analysis helps us navigate the complexities of macro-level resource allocations for risk-control, and helps us evaluate the costs and benefits of various macro-level interventions[428]". Let us consider his views and recommendations for the intermediate, operational level.

---

rules (and rent-seeking considerations), even though no health statistics backed this up (note: see chapter 1 on the roots of this classification in the Stalin era).

[424] See also Sparrow 2000.

[425] Location 285, Kindle edition.

[426] *Ibid.*

[427] Location 307, Kindle edition.

[428] Location 331, Kindle edition.

In Sparrow's perspective, "it is practitioners, not theorists, who need to know how to navigate the textured substructure of any general risk. It is they who have to know at what level of aggregation (…) to define a new project, and how many knots (harm-reduction projects) to take on at any one time. It is they who have to construct the data gathering practices and analytic lenses that enable them to spot the knots (risk concentration).[429]" His approach is one that squarely focuses on operational practice.

Some of the key *challenges* he identifies on the way to *harm-reduction* are linked to the *nature* and *characteristics* of specific harms: some "have a brain behind them" (involving authorities in a "game of intelligence and counter-intelligence"), some are "essentially invisible, with low rates of reporting or detection"), some correspond to powerful performance incentives, others are rare but potentially catastrophic[430]. These different challenges, to be properly addressed, require a set of methods and approaches that Sparrow covers in the second part of the book.

Another critical question is that of measurement – which, as we have seen above, is highly problematic. He points out the inherent tension between regulatory enforcement work that is mostly organized around "functions", "programs" or "processes" – and the need to give a "compelling account of *harm controlled*"[431]. Moving from reporting on *outputs* to reporting on *outcomes* can be mandated from above, but achieving it is far more difficult. It involves solving the questions of *causality* and *attribution,* as well as whether "it is possible to measure prevention", "accidents that didn't happen"[432]. Sparrow's contention is that it is in most cases practically impossible to prove causality, but that changing the way work is organized and performance is reported can allow to make a convincing case based on "the contributing micro-level outcomes: *the stories of the projects*[433]". While such a method will definitely not be "scientific" and will not "prove" causation, by building accumulated convincing micro-success stories, and as long as they "constitute significant progress towards important strategic objectives"[434], it will make it far easier for the organization to make a strong case for its effectiveness.

Without doubt, Sparrow's work is important, and it is influential among practitioners, because it focuses on a level of "operational challenges" that has generally been under-researched. His central recommendation in operational terms is to focus on "unpacking" aggregate harms and identify "knots", causalities, patterns, and structure interventions on this basis, i.e. by "projects" rather than through fixed functional structures. From our perspective, his emphasis on practice, and his suggestion that convincing patterns of effects may be more realistic than absolutely scientific attribution are clearly relevant. On the whole, however, we would not say that Sparrow's vision really is an "alternate account" of risk-based planning and risk proportionality in inspections and enforcement. Rather, it gives inspection officials very useful directions on how to make sense of problematic patterns, how to design more "creative" interventions.

Not only is our research focus at a somewhat more "aggregate" level, but we also believe that Sparrow's insights are best applicable within organizations that have *already* moved to a risk-based approach. His recommendations will then improve effectiveness, review organizational structures, put the question of risk (or "harm reduction") at the centre of operational decisions in practice and not just in theory. Thus, while we consider these insights as sufficiently important to cover them in some details here, we will make limited references to his work elsewhere in this research, as it mostly relates to the question of operational implementation within the context of an already "risk-focused" agency.

---

[429] Locations 366-375, Kindle edition.
[430] Locations 385-395, Kindle edition.
[431] Locations 2365-2372, Kindle edition.
[432] Locations 2422-2432, Kindle edition.
[433] Location 2516, Kindle edition.
[434] Location 2575, Kindle edition.

b. Dealing with risk: precaution, risk aversion, risk proportionality

i. Risk aversion and crisis-driven "panic" reactions

Introduction and definitions

In many areas of life, consciously or not, citizens rely on rules and regulations protecting them, and on these regulations being effectively complied with and enforced. Such expectation of protection underpins the trust in the food we eat, the products we buy, and the air we breathe. In practice, however, if designed inadequately or with unrealistic expectations, regulations can fail to work. In other cases, market incentives and contractual obligations may be sufficient, without the need for regulation to intervene. Often, implementation is the problem: insufficient guidance and support, or lack of resources for control and enforcement, or wrong methods, can all lead to disappointing levels of compliance. But there clearly remain "market failure" situations where regulations are indispensable to ensure safety and protect the public interest and where, if well designed and implemented, they can be very effective. Likewise, for some of these regulations, inspections and enforcement by state authorities are indispensable to promote compliance and, if done with the right methods, can ensure that regulatory goals are reached.

Over the past two decades, tools and methods of "better regulation" have been developed and put in practice, aimed at ensuring that existing and new regulations are of the efficient and effective kind. Somewhat more recently (but since at least 10 years), these improvement efforts have also extended to the whole "regulatory delivery" sphere, all the actions and tools that aim at turning regulation into practice, in particular regulatory inspections and enforcement. In spite of these tools and efforts, however, complaints abound that many new laws and regulations continue to be adopted that fail to pass muster in terms of necessity, cost-benefit and other key criteria, and political decisions on delivery tools and methods (licenses, permits, inspections, enforcement approaches) also frequently appear at odds with evidence and best practice, disproportionate, inefficient, or frankly counter-productive.

In some cases, this seems to happen because regulations, decisions, priorities are pushed through in response to sudden accidents, crisis situations, in a kind of panic reaction that has been called the "risk regulation reflex", a term coined by Margo Trappenburg in an essay she prepared for the "Day of Risk" conference, organized in May 2010 by the Dutch Risk and Responsibility programme[435]. The term "risk regulation reflex" is meant to refer to a mechanism leading to disproportionate government interventions surrounding a risk or following an incident. A corollary of the risk regulation reflex is that preventing, avoiding or compensating for risks is often seen as a government responsibility by default – in other words, the "risk regulation reflex" would be in some ways the *opposite* of "risk-based regulation"[436]. The "risk regulation reflex" concept can apply to both "short term" incident responses, and to the broader, "long term" trend towards ever more safety. It can designate "a trend towards ever more far-reaching safety measures which carry the chance of imbalance between the gain in safety and the costs and side effects of the measure, and the pitfall of public demand for a swift response following an incident leading to disproportionate measures" (van Tol 2012). From our

[435] Over 2008-2009, originally as part of the Netherlands' Inspection Reform Programme, increasing focus was put on exploring "overreaction to risk" and how to address it, building on the UK RRAC's work (van Tol 2012). This led in November 2009 to the creation of what came to be called the "Risk and Responsibility" programme (van Tol 2012, 2013) – in Dutch "Risico's en Verantwoordelijkheden".

[436] See Rothstein, Borraz and Huber 2013 on how the "duty of protection" (*"Schutzpflicht"*) embedded in German legal principles makes it difficult to implement risk-based regulation.

perspective, both aspects are essentially linked: disproportionate responses to incidents are made possible by a context of "risk aversion" and, in turn, successive incident responses end up building a trend. Both can lead to changes in inspections and enforcement practices that go opposite what "risk-based inspections" seek to achieve.

Is risk aversion on the rise? A disputed issue

Unfortunately, solid statistics on regulations, and in particular on how many may have been adopted as a result of such "reflex" situations, are hard to come by – meaning that it is impossible to prove beyond doubt that the phenomenon is real. Anecdotal evidence, as well as important studies[437], suggest however that risk aversion (e.g. in the form of the "risk regulation reflex") is a significant cause of inadequate policy responses – either directly (new rules developed in the immediate aftermath of the event), or by making their way into the election platform of a party, and being introduced after an election victory. In all cases, what happens is that political priorities trump analysis and evidence, and that these political priorities are defined based on risk avoidance and "absolute" statements ("this risk is unacceptable" and "this should never happen again").

In 2005 already, Tony Blair, then Prime Minister of the United Kingdom, issued the following warning: "In my view, we are in danger of having a wholly disproportionate attitude to the risks we should expect to see as a normal part of life. This is putting pressure on policy making [and] regulatory bodies (…) to act to eliminate risk in a way that is out of all proportion to the potential damage. The result is a plethora of rules, guidelines, responses to 'scandals' of one nature or another that ends up having utterly perverse consequences[438]." This same speech was quoted in *Rethinking Regulation,* a report published in January 2006 in Australia and summarizing the work of the "Taskforce on Reducing Regulatory Burdens on Business" (Banks 2006). This report opened with remarks on the growth of regulation, which covered Australia but could have been about many other countries: "Australia has experienced a dramatic rise in the volume and reach of regulation, in response to a variety of social, environmental and economic issues". It then moved on to discuss the possible causes of this regulatory inflation: "It is important to recognise the forces behind the growth in regulation if sustainable solutions are to be found. Perhaps the most fundamental of these is the changing needs and expectations of society itself. Some of this is a natural and desirable consequence of rising affluence and increased scientific knowledge. However, in the Taskforce's view, a more problematic influence has been increasing 'risk aversion' in many spheres of life. Regulation has come to be seen as a panacea for many of society's ills and as a means of protecting people from inherent risks of daily life. Any adverse event (…) is laid at government's door for a regulatory fix. The pressure on government to 'do something' is heightened by intense, if short-lived, media attention."

Both Tony Blair and the Banks report thus give a "classical" summary of the "risk regulation reflex": excessive reaction to adverse events, excessive demands for absolute safety and protection, resulting in regulations that go far beyond the needed and the reasonable. While the Banks report focused on regulations affecting businesses (and particularly small businesses, reminding that "regulatory burdens fall disproportionately on the economy's many small (including 'micro') businesses, which lack the resources to deal with them"), Tony Blair expounded also on the impact of such risk-averse regulations on "daily life": "something is seriously awry when teachers feel unable to take children on school trips, for fear of being sued" – and further in the same speech: "for example, one piece of research into a supposed link between autism and the MMR single jab, starts a scare that, despite the vast weight of evidence to the contrary, makes people believe a method of

[437] See e.g. Productivity Commission 2012 – page 316. All this work owes a lot to the work of the UK's Risk and Regulation Advisory Council – see RRAC 2009 (series of publications) in bibliography section.

[438] 'Common Sense Culture, Not Compensation Culture', Speech to the Institute of Public Policy Research, London, May 2005 - http://www.theguardian.com/politics/2005/may/26/speeches.media

vaccination used the world over is unsafe. The result is an increase in risk to our children's health under the very guise of limiting that risk". Indeed, the MMR vaccination scare is a perfect example of "scare" leading to adverse health effects. Problems with what used to be routine school activities (school trips, or bringing home-baked cakes) have also been reported (and felt sorely) in many European countries – though they do not always originate in new regulations, but sometimes in increased litigation and enforcement of liability originating from quite "old" regulations.

Critics have pointed out (Carroll 2006) that the Banks report was making important claims, but did not always have data to back them up. Showing that the volume of laws and regulations has increased may only reflect the calls for higher quality of rules and increased quality, and the estimates of administrative burden are (by the Banks report own admission) difficult to make and highly variable. Furthermore, again as per the Banks report itself: "While a number of studies have sought to estimate the economic costs of regulation in Australia, the limitations of such studies mean that the estimates should be treated with caution (…). Further, none of the studies measure the extent to which the compliance costs exceed what is necessary to achieve the policy goals underlying the regulations, which is the focus of this review. Quantifying this unnecessary element is even more difficult, and clearly". Indeed, it is difficult to convincingly prove (or disprove) that the regulatory burden has increased, and/or that regulation is ever more intrusive and covering areas of life that used to be freer, *and* doing so in ways that add little or no discernible safety or other benefit. It could conceivably be done by thorough analysis of changes in regulations, benchmarking across countries etc. – but it would require a significant research undertaking, and resources.

In short, there is some discussion as to whether such "risk adverse" responses are overall on the rise or not, whether the volume (and consequences) of poorly-designed policy responses they produce is increasing or not – and overall it is very difficult to quantify how large the effect of such policies is (see Helsloot, Schmidt 2012 and UK National Audit Office 2011). Available evidence however suggests that "risk aversion" and the "risk regulation reflex" are not insignificant problems – not only in economic terms, but also because excessive regulation undermines the legitimacy of public action, both because it hinders legitimate private activity, and because it fosters the illusion that the government can achieve "perfect safety", which is bound to be disappointed ("it can hinder society's self-reliance and resilience, restrict the freedom of citizens and businesses, diminish the government's authority as a result of promising too much" – van Tol 2012). In addition, a negative impact on the economy in turn will have significant negative impact on safety and health – as pointed out by Helsloot and Schmidt (2012): "life expectancy is strongly related to a person's income (…). Life expectancy actually increases up to seven years for people with a higher income compared to people which are poor, and the difference in the number of years the two groups experience a good health is as much as 16 to 19 years. A safer society, at least if we define safety in terms of average life expectancy, can consequently be reached by boosting prosperity in lower income groups" (see also Mackenbach, Kunst, Cavelaars 1997).

Thus in our view the limitations in evidence are not a major obstacle in terms of establishing the importance of risk-based approaches as a way to balance "risk aversion" trends. First, because "anecdotal" evidence of "regulatory creep" and "risk aversion" in regard with "daily life" activities is quite substantial, and the growing discontent it generates in a number of countries sufficient cause to think about how to alleviate it. Second, because there is also considerable evidence, through benchmarking in specific regulatory areas, that some countries within the EU, i.e. with many of the same fundamental parameters and many harmonized regulations, impose far more burdensome regulations and regulatory procedures (licensing, permitting, inspections etc.) than others – without additional safety to show for it in many cases[439] - and it is precisely this

[439] An important clarification is in order here: in *some* cases, countries impose higher regulatory requirements than is the case elsewhere in the EU *and* have a clear difference in results to show for it (e.g. several nordic countries in environmental matters). In

evidence that we intend to consider more closely in the third chapter. Finally, because in any case, regardless of overall trends in risk aversion or regulation, ensuring that the best possible policy decisions are taken in terms of effectiveness and efficiency is of public benefit.

In this perspective, rather than focusing the discussion on whether there is convincing proof of an increase in regulatory burden (which is debatable, particularly if we are talking about net burden, i.e. "burden less benefits"), or an increase in risk aversion in the society (with some clear examples in some areas, but also important counter-examples), the focus should be on what situations, contexts and systems produce bad decisions – and which ones can, on the contrary, foster good ones. To quote the authors of The Government of Risk: "macroscopic and world-historical perspectives on risk and its management may have their uses. But most of them do not explain, or even describe, variety within the putative 'regulatory state', 'risk society' or 'audit society'. Yet casual observation, academic inquiry, and official surveys alike indicate substantial variety in the way risks and hazards are handled by the state" (Hood, Rothstein, Baldwin 2001).

Understanding "risk regulation reflex" processes

One may wonder why reacting to a disaster would necessarily lead to the wrong response. Since the Middle Ages at least, if not earlier, regulations (and institutions) have come into existence in response to risks, real or perceived, and often in the immediate aftermath of disasters of some kind (be it a sudden event or a prolonged situation). This has been particularly true of the growing system of regulations and regulatory implementation structures that has developed over the past two centuries – covering occupational safety and health and labour rights, environmental protection, food safety etc. We have outlined some of this early history in our first chapter. While we attempted to show how much of the adoption of new rules and creation of new institutions was linked to risk *perceptions* (mediated by a number of social, political and economic factors), it is nonetheless clear that some important regulatory steps responded to very real risks. Just as clear is the fact that, in spite of the difficulties in causality and attribution, and the evidence that some improvements predated regulation, at least *some* of the improvements in safety and public welfare were driven by these regulatory changes.

Taking a couple of examples will help illustrate the point. In the UK, the 1833 Factories Act led to the creation of HM Factory Inspectorate in the same year, and the 1842 Mines Act to the creation of the Mines Inspectorate in 1843 (with increased powers from 1850). In both cases, this came in reaction to public opinion being shocked about working conditions in factories and mines (particularly for children and women). In the United States and much of Europe, as in the UK, mining accidents led to safety regulations being adopted, and often inspecting institutions set up, in the 19th century. The same goes for instance for the US Food and Drugs Administration, created in 1906 following scandals about adulterated or otherwise hazardous foods and drugs[440]. Tragedies caused by drugs touted as "safe" (e.g. Thalidomide) led to increasingly stringent prior approval regimes for medicines in the 20th century (and further scandals, such as the *Mediator* one in France, have led to further changes in these systems). Mid-20th century "killer fogs" in London led to pollution controls. The Seveso disaster gave its name to an EU directive (and its successive iterations), and other chemical disasters such as Bhopal in India, Love Canal in the US etc. all led to strengthened regulations and oversight[441].

---

such case, it becomes a question of cost-benefit analysis and of prioritization in values and objectives whether to opt for such stronger regulations or not. In other cases, countries impose considerable burden often through numerous permits, approvals etc., or additional regulatory norms (like the lift safety example we used above), with very little or no positive impact at all. This latter case is the one we are referring to here.

[440] http://www.fda.gov/AboutFDA/WhatWeDo/History/CentennialofFDA/default.htmUK

[441] See e.g. Balleisen, Bennear, Krawiec and Wiener (in press) as well as IRGC Conference presentations by the same authors

Even though critics of government regulations would argue that current occupational safety or food regulations impose too much burden on economic initiative, the part of these regulations that dates back to a century or more ago is widely accepted as having delivered considerable benefits at what appears to have been a very limited cost to economic growth, innovation etc. What, then, has changed so that nowadays dramatic events are said to lead far too often to regulatory responses whose costs outweigh the benefits they may bring (or even sometimes bring only negatives)? What is it that would explain a "risk regulation reflex" with overwhelmingly negative outcomes? The first change is probably the increasing marginal cost of adverting accidents and other hazards: the higher the existing safety level, the higher the cost of additional improvements in safety. As Helsloot and Schmidt (2012) put it: "every improvement curve flattens out at a certain point. Consequently, anyone who wants to achieve anything in the 'tail' of the curve needs to be very cautious about making substantial investments, as [their costs] can easily be disproportionate [to their benefits]".

This "flattening of the improvement curve" is a feature that is very difficult or impossible to affect through public policy – and thus there is an inherent character, to some extent, in the fact that further improvements in safety and health will, more or less inevitably, have greater costs than the ones that came from earlier "low hanging fruits". There are, however, a number of other factors that can lead to an excessively costly and poorly thought-through "risk averse" way of regulating, and they are often understood to be:

- Lower risk-tolerance, meaning that we tend to address issues that in earlier times would have been accepted as the normal state of things
- Difficulty for scientific evidence to overcome ideological preconceptions, pseudo-science, and fundamental psychological patterns with regard to risk
- "Positioning" of political and other actors (media, interest groups) in a world where information flows extremely quickly and where what used to be small, local news items swiftly become national or global. This leads to over-reaction, and to decisions being taken too quickly and without proper analysis, insufficient attention to regulatory design etc.

We would argue that all three points are important, and indeed there are factors pertaining to risk-tolerance and risk-aversion (and their psychological underpinnings), to the trust or lack thereof in scientific advice and in policymakers statements, and to policy actors – but the characterisation above leads to many misunderstandings of how unavoidable risk-aversion is (or is not).

Psychological aspects of the risk response

Indeed, psychological aspects are important, and indeed human heuristics are poorly suited to dealing with uncertainty and statistical aspects of risk (see Tversky and Kahneman 1974, 1979)[442] – but research and experience also show that, when engaging properly with the public, it is possible to discuss risk in a rational way and to ensure that risk perception does not necessarily degenerate into risk aversion, but rather that risk acceptance can be fostered. Indeed, while risk perception is essential in determining each member of the public's initial response to a risk or incident (see e.g. Slovic, Fischhoff, Lichtenstein 1982, Slovic 2000), what matters in the end is whether the initial perception is "frozen" or not.

Repeated research (see Helsloot, Schmidt 2012) has shown that, while simple questions asked without any background or any additional information tend to produce responses where people manifest strong risk aversion, this can change when additional information and context are provided. Indeed, people do not

[442] See p. 187 below for more discussion of human heuristics. For a specific discussion of availability heuristics and their effects on risk regulation, see Kuran and Sunstein 1999.

respond to risk only from a "more" or "less" risk perspective – but integrate a number of other values (fairness, equality, liberty, self-reliance etc.) (see Eeten, Bouder 2012[443]). Research suggests that "people seem to be able to make a difference between their own risk perception and what risks should be accepted reasoning from an administrator's point of view" – when given sufficient information on costs and benefits, they will balance the advantages of addressing a specific risk with its downsides and with other alternative uses of resources, whereas if only asked whether a given risk is important and worth addressing, they will usually answer "yes"[444].

The conclusion here would seem to be that the public may well be far smarter than usually given credit for – engaging members of the public takes time and resources, but can yield a far more balanced and rational approach to risk than relying on rushed "yes/no" questions with no context and information[445]. Balancing these findings and optimistic views on the possibility of a rational debate on risk comes other evidence that public discussions of risk are very difficult because of the problem of risk perception. We have already presented the findings by Slovic and others on how perceptions of risk are often very distant from what statistical estimates would suggest. In addition, other psychological factors mean that a discussion of a statistical risk may see the very salient and understandable risk ('death') be perceived far more strongly than the statistical probability ('one in a large number', which is very difficult to conceive). A more nuanced conclusion would thus be that attempting to have public discussions of risk is possible, but requires to set a discussion framework that starts in small settings and builds understanding of the issue (and of the data) among stakeholders (including the media). When risk discussions suddenly break out in public discourse without such an effort at building a joint understanding, the results tend to veer much more towards risk aversion and "panic" reactions.

Science, transparency, trust

Likewise, the public's relationship with science is also more complex than many experts would suggest, who mostly see the public as insufficiently listening to science and not able to properly distinguish "real" from "pseudo" science. Most of these conclusions lead their authors to recommend that efforts be made to ensure that the public defers more to scientific advice, but in ways that seem more like "communication" and "propaganda" than real engagement.

There is certainly a share of the public who will not accept scientific findings and rather adhere to other views – be they based on religion, ideology, conspiracy theories or any other worldview. When a significant share of the population holds such views, it is important to acknowledge them in the public discussion, including indicating that the policy decision will *not* be based on them, but on scientific findings and utility maximization[446]. What matters more to us here is that, for those members of the public (typically, the majority) that do *not* hold deeply views that are fundamentally at odds with a scientific perspective, trust in scientific advice (and in policies that claim to be based on it) can be built up – and can be destroyed as well.

Dissimulation or manipulation of evidence, claims of full harmlessness for things that later are proven to have been extremely hazardous (or the opposite: claims that something is very dangerous whereas further evidence

---

[443] For an excellent overview of the different values that can underpin radically different approaches to risk and trade-offs in the criminal justice field see Buruma 2004.

[444] In one of the experiments presented by Helsloot and Schmidt (2012), 35% of respondents essentially change their mind within the course of one single interview, when moving from simple dual questions to a more considered discussion and asked to put themselves "in the shoe" of a policy maker.

[445] Supporting this view, see Posner 1998 and Esptein 2008.

[446] Discarding them without even a proper mention, by contrast, decreases legitimacy by making the process "unfair" from a procedural justice perspective, as dissenting views are not even given a "voice" (regardless of the final policy outcome).

demonstrates it to be less so), can severely damage the public's trust in science – or at least in official claims for policies to be based on "science". We will discuss these issues in more details further in this research.

Policy actors – the risk aversion cycle, and the problem of "risk experts"

Finally, the question of policy actors is important, and has at least two aspects – one being the way all actors in the "risk regulation reflex" are linked through a kind of cycle, the other being the activities and impact of "policy entrepreneurs".

The circular aspect is the way, in a "risk regulation reflex" process, all actors in a way attribute responsibility for decisions and actions to someone else: the media claims that the public is outraged and demands action, politicians say they have to act because the issue is all over the media – and civil servants claim they are compelled to act by politicians and the media. As for the public, it faces a barrage of media coverage, and politicians all promising that "it should never happen again", and feels reinforced in all feelings of risk aversion. This relationship has been called "Februari's Circle" (van Tol 2014) after Maxim Februari, who exposed it as part of the work done for the Risk and Responsibility programme (van Eeten *et al.* 2011). The crucial element of this circle is that no one is taking responsibility – and everyone claims to be doing their job. The media say they have a responsibility to voice public concerns (and an interest in "crisis", which sells well). Politicians say they have to respond to their constituents' demands (and an interest in winning, not losing, elections). Civil servants say they have a duty to follow priorities laid out by elected politicians (and an interest in keeping their jobs). In all this, interest is more evident than duty – and the attitude of members of the public is typical of the "Not In My Backyard" (NIMBY) pattern (Helsloot, Schmidt 2012).

As we have indicated above, this circle is not a fatality: breaking may be possible, by providing the public with more information and context, and initiating a real public conversation about the risk at hand. This requires, however, initiative from at least one group of actors. This is not easy – as Carrigan and Coglianese (2012) put it: "Intense reactions by the public (…) drive an intense desire by politicians to take action. Under such circumstances, taking any action targeted at the regulatory process, regardless of how well or poorly crafted, will be better politically than taking no action at all. Political incentives point in the direction of quick legislative action that responds to calamities. Voters focus much less on considerations of how a law will be implemented than on the enactment of a new law itself (Mayhew 1974; Mazmanian and Sabatier 1983). Legislators can reap rewards from passing legislation regardless of whether doing so turns out to be realistic or effectual". However, we have seen that research also shows that engagement with the public can yield real changes – thus, if the "circle" can be interrupted, the pressure to act regardless of effectiveness will stop. This report aims at presenting ways in which space for such a "rational conversation" can be created.

In addition to actors in the "circle" seeking to push responsibility on others, there are some specific actors who *actively* seek to strengthen risk aversion, who have an active interest in reinforcing the reflex, in making the particular risk appear as particularly serious so as to maximize the response. What many authors call "policy entrepreneurs" can be of many kinds, and have been studied from a variety of angles (see e.g. Roberts, King 1991 – Mintrom, Norman 2009 – Cohen 2011). The importance of "policy entrepreneurs" as one of the elements shaping response to risk has been pointed out by Hood, Rothstein and Baldwin (2001)[447], and the presence and activity of these "entrepreneurs" is for them one of the elements that can lead to different "risk regulation regimes". From the perspective of the RRR, which represents a specific case of "risk regulation regime" (one with particularly strong response compared to what could be expected from a rational analysis of the "market failure" – see again Hood, Rothstein, Baldwin 2001 for a broader typology), a feature seems to

[447] And we have seen that this role is not new, as it also had its importance in the creation of the US FDA in the early 20th century.

be that there are "policy entrepreneurs" particularly successful at pushing for such a response. These "policy entrepreneurs" were generally already pushing for their favourite policy, and the incident gives them an opening: "Crises provide opportunities for policy entrepreneurs to place at center stage those solutions they have already been seeking to see adopted (Kingdon 1984:91). Even if those solutions were not developed to address the particular problem at hand, politicians often feel compelled to consider them— to "do something" (Carrigan, Coglianese 2012).

They can belong to different categories – private businesses in some cases (e.g. suppliers of equipment or services to address the particular risk considered, e.g. lifts retrofitting as in the French example presented above), NGOs in others (e.g. those focusing on environmental protection, or some trade unions etc.), but also "experts" (independent or affiliated with consulting firms, research institutions, NGOs, businesses etc.) – and they are also quite often *inside* public administration (and in such cases, pushing for more regulation in their sphere of competence is a way of entrenching their importance, and their budgets – see Helsloot, Schmidt 2012).

Not all such "policy activism" is motivated by self-interest, far from it – "risk experts tend to really believe, and policy makers are made to believe, that an incident is proof that regulation should be tightened" (Helsloot, Schmidt 2012). The difficulty for civil servants and elected officials alike (and for journalists) is to decide whether these "risk experts" are right – to screen their proposals, or to review existing rules adopted in a previous "RRR moment". Indeed, "knowledge is required in order to determine what rules are disproportionate and can therefore be repealed. This knowledge is usually only available to the risk professionals of policy departments and their external advisors" (*ibid.*).

Thus, again, an essential step in order to avoid risk-averse, "reflex-driven" decisions is to provide time and space for careful consideration of arguments and evidence, rather than relying immediately on whichever "solutions" are advocated by "experts" which, even in the absence of material interests, will have a personal investment in their own field of study and expertise.

Modelling the policy decisions in a risk context

Another, more detailed way to look at these factors of "reflex" reactions and their consequences is the model proposed by, Balleisen, Bennear, Krawiec and Wiener (in press). In this model, crisis events can lead to small or large changes in risk perceptions, and the latter again to major or minor shifts in policy agenda. The magnitude of changes depends to a large extent on how the crisis fits or contrasts with baseline risk assumptions, and how the perception of the crisis is mediated by ideologies, heuristic models, narratives ("master-stories") etc. The interplay of interest groups' agendas, resources available, trust or distrust in specific institutions or actors, etc. then again influences whether the changes are substantial or mostly "cosmetic". Weber (in Balleisen *et al.,* in press) adds a psychological dimension to the analysis (based in particular on Tversky and Kahneman): for instance, humans tend to under-estimate the actual risk of events that are common and that they perceive as "normal", and to over-estimate the risk of events that have a very low probability but that they have previously experienced. There are many psychological mechanisms which mean that perception of risks by non-experts (be they politicians, journalists, citizens) can differ widely from what data shows the actual risk level to be. This is of course one of the primary reasons why over- (or under-) reaction to accidents and crises can occur.

In terms of sequence of events and reactions, this model sees events as being first mediated through baseline risk assumptions, and then modulated by a series of filters (ideologies, "master-stories", heuristics, media) in order to produce a "causal narrative" of the crisis. Depending on the different aspects of the context, this may result in blaming culprits or scapegoats, looking at structural issues, "policy regret" or bias confirmation. The

causal narrative may be agreed upon, or disputed. Then, the causal narrative or narratives themselves get complemented by expert analysis (or analyses) and the whole agenda or "policy menu" gets itself filtered by interests at play, resources available, institutional structures and the level of trust (or distrust) in institutions), to result in policy decisions. Depending again on the whole set of events and context, these may be "cosmetic" or "substantial" changes.

In a more formalized way, this model emphasizes the same factors as the "risk regulatory regime" approach of Hood, Rothstein and Baldwin (2001) or the key elements of the RRR evidenced by van Tol (2012) and Helsloot and Schmidt (2012): the importance of a context where values and visions of the public and the different actors shape how they perceive and react to risk, the impact of the intervention of experts and other actors to shape events into a "causal narrative" and a policy agenda, etc. It adds to this the importance of institutional capacity (or lack thereof) in steering the final policy decisions – Helsloot and Schmidt (2012) present, however, several examples of how the RRR can lead to policy decisions in favour of new regulations even in the absence of capacity to implement them (in several of these cases, the new regulations later end up being abolished, because they have not been seriously implemented).

All these analyses and models concur in highlighting the importance of *perceptions* and *shaping* of the issues, and also of what interests are at play, and what context the crisis occurs in. The key about the "reflex" mechanisms is the tendency to react *too fast* to the event – without giving sufficient time for inquiry and analysis. Against this, Bennear (in Balleisen *et al.,* in press) suggests that the answer should be "deflect" (take visible but inconsequential actions showing political attention but not locking-in potentially harmful decisions – thus giving time for further consideration) or "reflect". The key seems to be to create a shared understanding of this need to defer meaningful action until the situation has been more fully understood, to create the "conditions of possibility" for this time and "breathing space".

Relevance to the inspections and enforcement issue

The trends, research and discussions we have attempted to summarize relate to "regulation" in general, and not only or specifically to inspections and enforcement. They are, however, fully *applicable* to the inspections and enforcement "stage" of regulation. As mentioned above, the Netherlands' Risk and Responsibility programme, which led to the definition of the "Risk Regulation Reflex" concept, itself originated from the Netherlands' Inspection Reform programme. This inherent link between reaction to risks and regulatory control and supervision is an important angle for our study, and one of the areas where addressing risk aversion is most important.

Indeed, when incidents happen, inspectors and inspection services are often among the first to be blamed – and stricter, more frequent inspections very often top the list of "risk regulation reflex"-driven requests. When new technologies or practices emerge, inspectors may be the first to notice them – and possibility in some cases to prohibit them. Inspectors are on the "frontlines" of regulation, the main interface between rules and those who have to abide by them (mostly businesses, but also citizens).

Most of the difficulties related to inspections and enforcement in a perspective of rational risk management and risk mitigation come from a number of fundamental misconceptions on inspections themselves (their role and methods), and on compliance and safety (and their drivers) – misconceptions that are not only held by many members of the public (as well as "experts", interest groups etc.) but also by a number of inspectors and inspectorates managers.

These misconceptions revolve around the assumption that more inspections and stricter inspections (or more and stricter control, police checks etc.) will mechanically drive higher compliance, and that this will in turn automatically result in higher safety. This assumption in turn stems from a vision of compliance with rules is

primarily or exclusively driven by deterrence, fear and rational calculations. It also implies a belief that inspections, checks etc. do not have significant adverse and unintended effects. In turn, this excessive and unfounded assumption that deterrence is the major driver of compliance (and safety) and that inspections and checks are thus the primary tool to be used (and used as much as possible) fosters excessive expectations from inspections – i.e. that they should manage to ensure perfect safety, complete protection from risks in a given field.

Thus, the consequences of risk aversion on inspections and enforcement questions are serious. Understanding better the mechanisms of the "risk regulation reflex", and the ways to achieve a more balanced approach to risk, are essential to provide a foundation for "risk proportional" inspections.

ii. Uncertainty, trade-offs and transparency

Conflicting goals – and the pitfalls of excessive certainty

Individuals, societies, governments, international or "supra-national" organizations, all have sets of goals and objectives that coexist but may in some cases (or even frequently) come in conflict. Many would argue that the quest for more material well-being (observed both at individual and social level, and backed up by policies supporting economic development and the private sector) can conflict with another objective of both individuals, societies and public bodies, the protection of health and more broadly the environment. Certainly, it is not always the case that these goals conflict, as for instance the whole "green growth" idea (and realities) show. But there definitely are instances when objectives (and the values underpinning them) conflict. This conflict is clearly visible about risk – with risk-averse, "precautionary" demands on one side, and the push for a more risk-proportional, freedom-enhancing approach on the other.

A very good example of such conflict in goals, and of its possible consequences in terms of regulation, is presented by Ragnar Löfstedt in his article on the 'Swing of the Regulator Pendulum' (Löfstedt 2004): "the issue of both improving and implementing regulations are closely linked to the three main drivers of EU regulatory concerns: competitiveness, good governance and sustainable development. For example, if regulations are not improved, not only will European competitiveness be adversely affected, but also the criteria for good governance will not be met. Similarly, if environmental and health regulations are not properly implemented how can the EU state that it is taking sustainable development seriously?" He goes on to indicate that "the three drivers (competitiveness, sustainable development and governance) are, according to the Commission, closely interrelated and compatible. The Commission has long held the view that there is no actual conflict between environmental protection and competitiveness. It stated in the 1993 5th Environmental Action programme that: *The perceived conflict between environmental protection and economic competitiveness stems from a narrow view of the sources of prosperity and static view of competition*." While not commenting on this optimistic view held by the Commission, Löfstedt further exposes the tensions between the "precautionary" and "impact assessment" philosophies, and suggests that, in attempting to build credibility by showing "fairness" through "tough" decisions against business interests, the EU regulatory bodies have probably overshot their target and that the pendulum is likely to start swinging back towards "risk assessment" rather than "harm prevention".

This example suggests implicitly that there are, indeed, trade-offs – at least, in the author's perspective, between legitimacy of public authorities and economic growth. But we would argue that the cases presented in the article actually show that there is a tension between environmental and health protection and, if not economic growth overall (on which it is more difficult to comment because of the complexity of the effects involved), at least the availability of cheap products on the market, and possibly short-term job creation. One

of the examples used by Löfstedt is the ban on virginiamycin in animal feed, and the use of the precautionary principle by the European Court of First Instance in its 2002 ruling against Pfizer. Since the article states that "there was no reputable scientific evidence that there was a transfer of antibiotic resistance to humans as a result of the use of the antibiotic in animal feed" and further suggests that the decision was excessive (and an example of steps that may in the end trigger a "swing of the pendulum" in the other direction), it is worth looking (of course with the benefit of hindsight) at how well this decision has stood the test of time in terms of science and risk assessment. In its latest guidance for industry on the subject of use of antibiotics in animal feed[448], the United States Food and Drugs Administration (FDA) emphasizes the need "to help phase out the use of medically important antimicrobials in food animals for production purposes[449]". In 2013, the US Center for Disease Control (CDC) stated in its *Antibiotic Resistance Threats in the United States* report: "Antibiotics are widely used in food-producing animals (…) This use contributes to the emergence of antibiotic-resistant bacteria in food-producing animals [which] are of particular concern because these animals serve as carriers. Resistant bacteria can contaminate the foods that come from those animals, and people who consume these foods can develop antibiotic-resistant infections. (…) Scientists around the world have provided strong evidence that antibiotic use in food-producing animals can harm public health (…) Because of the link between antibiotic use in food-producing animals and the occurrence of antibiotic-resistant infections in humans, antibiotics should be used in food-producing animals only under veterinary oversight and only to manage and treat infectious diseases, not to promote growth."

We have quoted on purpose from US rather than EU agencies, because many authors (see e.g. Löfstedt 2004, Wiener 2003) would agree that they have been (at least in recent decades) rather *less* precautionary, and because (partly as a consequence of this regulatory stance and partly as a result of different economic structures) antibiotic use in animal feed is considerably more widespread in the US than in the EU. The fact that the FDA guidance documents are voluntary (a clear result of the need to balance safety issues and economic interests, and of the difficulty to overcome industry resistance) cannot obscured the fact that both the FDA and CDC are highly concerned and are trying hard to eliminate the routine use of antibiotics in animal feed, particularly when there is no disease being controlled and antibiotics just function as growth aid. In April 2014, the FDA released a list of "voluntary withdrawal" including 16 Antimicrobials for use in food-producing animals[450] – it included virginiamycin, the drug at issue in the *Pfizer* 2002 case. It seems that the Court's "precaution" was not so mistaken and groundless after all.

This shows the importance of caution when considering risks where significant uncertainty exists and knowledge is still under development. While designing adequately proportionate decisions in cases of well-known and understood risks is in general possible, there is a strong case to be made for a combination of "precaution" and "proportionality" when dealing with uncertainty. This may occasionally result in decisions that hindsight shows to have been excessive, but also in a number of other cases may result in avoiding very significant damage or disasters (see European Environment Agency 2001 for numerous examples). This is true not only for strictly-speaking "regulatory" decisions (adoption of new rules) but also for inspections and enforcement decisions. Inspectorates are often expected by public opinion to immediately address any risk, even when that risk is not certain, through control visits, withdrawal of products, sanctions etc. In cases where the "uncertain risk" is covered by the agency's mandate (i.e. it has authority to act), but also in other cases

[448] Guidances #209 issued April 13, 2012 - #213 issued December 2013 – both referencing guidance #152 issued October 23, 2003 – see http://www.fda.gov/downloads/AnimalVeterinary/GuidanceComplianceEnforcement/GuidanceforIndustry/UCM299624.pdf and http://www.fda.gov/downloads/AnimalVeterinary/GuidanceComplianceEnforcement/GuidanceforIndustry/UCM216936.pdf

[449] "FDA's Strategy on Antimicrobial Resistance - Questions and Answers" http://www.fda.gov/AnimalVeterinary/GuidanceComplianceEnforcement/GuidanceforIndustry/ucm216939.htm - see also http://www.fda.gov/AnimalVeterinary/SafetyHealth/AntimicrobialResistance/JudiciousUseofAntimicrobials/default.htm

[450] http://www.fda.gov/AnimalVeterinary/NewsEvents/CVMUpdates/ucm392461.htm

(because the agency could be at least lobbying to have its mandate extended), it is crucial for them to have a transparent approach to how they seek to balance risk management and precaution.

This applies for instance to authorizations and supervision of the use of medical drugs and devices, chemicals (food additives, pesticides etc.), or any other new technology. When considering the long-run, such a careful balancing act is important, and adopting a 'risk-based approach' should not be understood to automatically mean discounting risks that have not been proven significant simply because data is still lacking (as opposed to well-known risks where the data clearly points to their being of low importance). Indeed, a track record of discounting risks when uncertainty is significant, and of subsequent damages where it had been claimed that there was none to be feared, results in undermining the credibility and legitimacy of public authorities and their scientific advisors – and thus in undermining support for risk-proportionality. It is not just a question of costs and benefits in terms of life and health, and economic and social impact, but of the "snowball" effect that credibility loss will have.

Understanding and accepting trade-offs

A far better path towards understanding the "risk regulation reflex" problem and laying out potential solutions seems to us to be sketched out in the contribution of Cary Coglianese and Christopher Carrigan to the collective volume *Regulatory Breakdown*. Quoting them: "Is it possible that the ultimate failure of the U.S. regulatory system is that the American public, through its elected representatives, asks regulators to oversee activities that are at once desired but also deadly?" (Carrigan, Coglianese 2012). In other words, there *are* trade-offs: to a certain extent, different goals may be compatible, but at some point, they may conflict with each other, and choices (conscious or unconscious, open or hidden) will have to take place.

This is important in terms of managing expectations from inspections and enforcement agencies specifically, and not only expectations from government regulation generally – and thus in achieving support for risk-based approaches that are *explicitly* founded on the premise that preventing every risk is impossible, and that there are some risks where the costs of attempting prevention would be higher than the potential benefits. Thus, the very idea of trade-off is central to risk-based inspections, and the refusal of trade-offs is a key driver of risk-averse approaches, and of attempts to inspect every establishment, and to practise "zero-tolerance" enforcement.

As Carrigan and Coglianese point out, denying these trade-offs may well be one of the key reasons behind the RRR – as "insufficient" or "failing" regulation becomes an ideal scapegoat when something goes wrong. Quoting them (*ibid.*): "Calamities, we suggest, bring with them strong tendencies for faulty assessments of both underlying causes and necessary reforms. These tendencies are due to a host of factors, including both psychological biases as well as nuances in the policy process itself. The pressure politicians feel to adopt change even without solid policy analysis (…) means that solutions can end up being adopted that are either unrelated to the true cause of disasters or that actually work at cross-purposes to improving conditions. In addition, sometimes the underlying problem may not have to do with the (…) operations of the regulator or the regulated industry but may instead reflect inherent societal choices about trade-offs."

Disasters easily lend themselves to faulty assessments, based on heuristics that humans have developed to survive in their natural environment hundreds of thousands years ago, but are increasingly inappropriate to understanding situations in a technologically advanced environment and highly complex societies (Bennear 2014 and Carrigan, Coglianese 2012). Again quoting from the latter: "psychological and behavioral economics research (…) support the notion that people tend to focus more on worst-case outcomes and to believe that vivid events are more common than they really are (Tversky and Kahneman 1973). Moreover, researchers studying these phenomena— known as the "availability heuristic," along with other cognitive biases— also

report that they can be exacerbated by the media, which for obvious reasons tend to focus on especially dramatic events (Shrum 2002)." In such situations, regulators and regulations provide ideal points of fixation for negative emotions. The "culprits" in the narrow sense may be the business operators or individuals who were directly involved in the disaster, but regulators often end up receiving nearly as much blame. They form ideal "scapegoats" to blame for something that went wrong – regardless of whether this was in fact at all possible to predict, whether there were any structural elements or not.

If indeed the issue is fundamentally linked to the refusal to confront contradictions inherent to multiple goals, and to accept trade-offs, then "scapegoating" regulators and calling for stronger rules and enforcement is a way to continue this refusal. It is convenient for politicians, who avoid confronting their own failures (see for instance the case of the *Deepwater Horizon* in Carrigan 2013), and in a way for citizens as well, who do not have to make hard choices (at least not consciously). Achieving a more risk proportionate, approach to regulation and regulatory enforcement would thus start by making tensions and contradictions between different goals and aspirations clear and visible. From this, a rational conversation could be had regarding the potential trade-offs, the possible ways to reconcile conflicting goals to some extent, and the limits of this. On this basis, rational policy decisions can then be taken, with a clear view of what upsides and downsides they entail.

The uses and limitations of science in relation to risk and regulation

Risk-based regulation aims to rely on evidence and data in order to assess risks and decide on the adequate response, and this applies to risk-based inspections of course as well. In most areas, assessing risks in a "non-subjective" way requires the use of scientific findings – but this is not always as easy as many would think it is, because science is complex, incorporates uncertainty, and cannot answer all questions (and in particular cannot answer values-based questions).

A cursory review of developed countries in particular (but even many emerging economies) will easily show that "scientific advice is found almost everywhere in our technological cultures" and that, for many scientific advisory bodies, "the emphasis is on translating the state of scientific knowledge to make it useful for politics and for policy making" (Bijker, Bal, Hendriks 2009). Even though some of the institutions involved in scientific advice go back a very long way in time (like the Netherlands' Health Council, the *Gezondheidsraad*, which was founded in 1902), there does appear to have been an increase in the reliance on scientific advice in public policy, or at least the push for increased reliance, in the past three of four decades. This can be linked at least in part to major incidents – as a way to react to these not in a "reflex" way, but by improving the adequacy of policies and regulations in particular, through the incorporation of the "best available" science. In the case of the EU, around the mid-90s "amid scandals over industrial safety (Seveso), 'mad cow disease', dioxin contaminated food and oil vessels safety, the EU reconsidered the role that scientific evidence could and should play in its decision-making system" (Alemanno 2014). More broadly, the increasing emphasis on scientific advice in policy making can be tied to the increasing complexity of technologies employed both by businesses and in the private sphere, and the need to take decisions in front of issues where prior experience or a decent education are clearly insufficient guidance.

The increased reliance on science, part of the broader trend towards more "evidence based" policy making (of which RIA is a particularly characteristic example), is not only the result of technological change, however – and it is also not fully uncontroversial. On a fundamental level, one can argue that founding policy decisions exclusively or primarily on scientific evidence is in itself a major policy choice, reflecting a utilitarian ideology, and not (as it is often presented) a "neutral", "non-ideological" approach. Very often, in fact, "on contested topics (…) science, values and politics collide". The "utilitarian" perspective, which would have science be the primary guide for policy choices, and statistically predicted impact on human life the key indicator, has been

vehemently criticized from many corners (from, say, the religious right to the radical left) as reductive and as ignoring the role of "higher" (or at least "other") values in policy choices (for an early example of such criticism, see e.g. Slama 1993). The reason it is essential to remind of this here is that science can in any case not give the answer as to "what should be the right policy" – it can only, at best, indicate *which instruments and specific norms* are likely to be most appropriate *for given policy parameters*. For instance, if safety and health are the policy priorities, smoking bans and all measures against smoking will be welcome. But if individual freedom of choice is considered a higher value, then such bans and policies will be opposed (see Slama 1993). The only things science can say are (a) what the impact of smoking on health is (medicine and biology) as well as, to some extent, (b) what measures and tools are more likely to lead to reduced smoking (behavioural science, psychology, socio-legal studies etc.).

In addition to this fundamental limitation, there are many situations (and indeed, often in the "hottest" topics) where science is simply uncertain. Of course, at its heart, science always includes an element of uncertainty, in the sense that a better understanding of reality may always emerge – but "stronger" uncertainty is what matters here, that which is at stake in issues which are still only imperfectly understood, and where as a result diametrically opposing viewpoints can both claim to be based on "science" (as in the Endocrine Disrupting Chemicals – EDCs – "controversy", even though the vast majority of scientists appear to be on one side, i.e. the one that points out the hazards of EDCs[451]).

To summarize, there are several fundamental, intrinsic limitations to what "answers" science can give to public policy issues:

- Science cannot address conflicts between values, or respond which values to prioritize
- When a policy choice is likely to have conflicting impacts on different aspects or indicators, science cannot answer on which one should be given priority
- In fields where important uncertainty remain, it can only give answers which are affected by this uncertainty, i.e. based on probabilities
- Thus in all cases science cannot make choices – scientific advice can, rather, be a "honest broker" or "cartographer" that "helps decision makers to choose wisely between the available options" or at least understanding the implications of different "policy paths" (Wilsdon 2014).

The specific case of "scientific uncertainty" – dealing with uncertainty, dealing with risk, two different but connected problems

In many situations where regulators are under pressure to act, but also subject to criticism for over-reacting, science is in fact not fully clear. Whereas scientific issues are not in debate for instance in the *Deepwater Horizon* disaster or in the Foot-and-Mouth crisis (and the questions are only about the proper tools to address technological or epidemiological issues, and trust deficits), they are or were very much openly debated or at least "not fully solved" in cases like the BSE ("Mad Cow") crisis, or EDCs and the right response to give them.

There is, indeed, a tendency (on many sides) to present scientific opinion or advice as "one" – and to see problems only in terms of ensuring that scientific evidence gets accepted and acted upon. Quoting an influential report on *Enhancing the role of science in the decision-making of the European Union,* for instance (Ballantine 2005), the only limitations it sees to scientific evidence are "policy-makers and decision-makers [being] often unable to make use of scientific advice", "lack of public confidence in the utility of scientific evidence, particularly in managing risks to human health, which limits its effectiveness", "difficulties in

[451] The controversies on GMOs would of course be another example, but their complexity and the passions at stake are even greater, and in addition the "scientific arguments" used by both sides tend to show that they (on purpose or not) do not even speak about the same issues – many proponents of allowing GMOs cultivation and sale emphasize studies showing innocuity on human health, but many GMO opponents do not focus on human health effects but rather on the environmental impact.

obtaining 'independent' and 'excellent' scientific advice" and the fact that "some influential groups[452] do not accept that scientific evidence is an appropriate input". We contend here that this is an exceedingly restrictive and "technocratic" view, that assumes the answer is clear and beyond doubt, and the only problem are "people" and "politicians" not listening or unable to act upon scientific advice. The reality is far more complex.

Scepticism is often grounded in major failures in the past

If many people (or "groups") show limited trust in what is presented to them as being the state of science, it can be not only because they conflict with their values or "ideologies", but because past experience has shown the limits of claims of safety of technologies based on "science" [regardless of whether or not the claims were indeed based on science or just presented as such].

Chemicals or drugs later found to be highly toxic (and remaining actively toxic for extended periods) remained in some cases on the markets for decades – with both instances of their toxicity having long been known, or of their being originally seen as safe and knowledge of their toxicity only gradually emerging. Infamous cases that have made history in the worst way include thalidomide, which was marketed as perfectly safe for several years in a number of countries, and led to around 10,000 birth defects leading to infant deaths and phocomelia. Diethylstilbestrol likewise was prescribed for three decades to pregnant women in the mistaken belief it would reduce the risk of pregnancy complications and losses – and not only had no positive health effects, but led to cause a variety of significant adverse medical complications during the lifetimes of those exposed (in particular genital tract diseases, e.g. vaginal tumours and uterine malformations). PCBs and other chlorinated hydrocarbons were recognized early as toxic due to a variety of industrial incidents, but serious regulation was only introduced nearly forty years after the first studies, in the 1970s. DDT was used for decades before serious attention was given to its adverse effects, which had been hitherto noticed only by a few scientists. Significant campaigning against the massive use of this chemical only started in the early 1960s, after several decades of massive use worldwide. Asbestos and lead, two naturally occurring chemicals, had harmful effects on health that were known in part since ancient times (at least for lead), but serious regulation of their production and use took often decades to be imposed (with the United States only banning lead-based paints in 1971, Europe lagging at least a decade after the US to ban lead in gasoline etc.) – industry associations during this whole time made considerable efforts to resist regulations and try and discredit scientific expertise that showed the hazards caused by these materials.

We have chosen these few examples on purpose, as particularly well known. They have in common massive adverse effects, and the fact that they were marketed as perfectly safe and warranting little or no precaution (thalidomide and diethylstilbestrol were indeed specifically targeted as pregnant women, the most vulnerable population of all). In some cases, active dissimulation was involved – adverse effects were well known and hidden. In others, adverse effects were not really known, but no efforts were made to investigate whether the compound was really safe, and it was intensively marketed as such. They should remind us that, when individual citizens, NGOs or indeed scientists are sceptical about claims of innocuity, they are not refusing "scientific advice" (as Ballantine and others would put it) but showing legitimate caution in front of statements that probably overstate the confidence we should really have in many products' harmlessness. Being sure of the (absolute or relative) harmlessness of chemical compounds that are novel and are being put into massive production is extremely difficult, if not impossible, at least in a short timeframe. Deciding between a precautionary stance and a more "growth oriented" one is a matter of balancing risks, opportunities, and uncertainty – it is a matter on which a rational conversation can be had, and rational people on both sides can

[452] Given the make-up of the Steering Group for this report, with many industry representatives, the "groups" are clearly meant mostly to refer to NGOs – but could also be understood more broadly.

disagree. It is not a topic where a simple "scientific truth" can be told and any disagreement should be seen as baseless obscurantism.

Openness and transparency are indispensable to build trust

Using "science" as a foundation for risk-based regulation, and specifically risk-based inspections and enforcement, is thus not a simple matter of following "science" as if it were just one clear set of directives. We would argue that the first step is building real trust through transparency, including transparency about uncertainties and disagreements. Not paying attention to uncertainties, full transparency and the need to clearly show divergences of opinions may have been one of the causes for the controversy surrounding the former Chief Scientific Advisor (CSA) to the President of the European Commission, whose office was not renewed under the new Commission: "if her mission is to strengthen the role of science within the policy process, it is manifest that the CSA cannot and should not do that alone. It is only by rendering public a possible divergence between her advice and the political decision that the CSA's ontological mission to promote science in government could be accomplished. Of course, this is not to suggest that scientific input should prime over other sources of advice, but that when a tension exists between the two this should be rendered public" (Alemanno 2014)[453].

If scientific advice is to be any use in making the public trust the risk-based approaches of inspectorates, and their claims to an adequate balance of costs and benefits, the scientific advice itself needs to be trusted. However, in some instances, this trust has been harmed considerably by prior experience (see above), and by what is seen as attempts to push policy decisions that result from choices and prioritizations as "the only choice". Transparency is needed on what are the uncertainties, the options and the costs associated with each one. Scientific advice should not mean advocating only one policy option, at least in many or most cases, but rather laying out clearly the upsides and downsides of different options. When significant uncertainty is involved, different scenarios should be sketched out, the costs of different options clearly presented, as well as their potential benefits.

If we take an issue like EDCs, simply stating that their risk to human health is "hypothetical at best, possibly illusory, and certainly never scientifically established[454]" appears to be an overstatement that is damaging to the cause being advocated, because in front of the evidence already collected (WHO UNEP 2012, which comes on top of 10 years of research after the first 2002 report), this appears at best as an overstatement, at worst like as fully misleading. It does not ensue that the decision should be an "outright ban" (which Julie Girling is advocating against) – but certainly the policy debate cannot simply be dismissed by trying to disparage or dismiss the findings of what appears to be the clear majority of scientists specialized in this field.

In conclusion, while scientific advice is an indispensable element of proper risk-based approaches to inspections and enforcement, it can in no way provide the sole source of rules, decisions, guidelines and practices. It is essential to understand and acknowledge that political decisions will be needed, based on values – as well as "technical" decisions by inspection officials, based also on values, combined with experience and a variety of heuristics. Combining a form of "precautionary principle" with a risk-based approach to inspections is not necessarily a contradiction. Precaution can be understood as a tool to use in the face of uncertainty, and it would not be impossible for a regulator to decide to be precautionary in the face of risks that cannot be assessed with certainty, but to otherwise make its approach proportional to risks in terms of requirements and enforcement decisions, and targeted on risks in terms of resources. The precautionary principle would just be in this way a heuristic tool to assign a rating to hazards that are subject to high uncertainty. A

[453] See Blanc, Macrae and Ottimofiore 2015 p. 59 for a summary on this controversy.
[454] Julie Girling in the *Wall Street Journal* on 23 January 2014:
http://www.wsj.com/articles/SB10001424052702303947904579336611208924306

"precautionary" regulator would give a higher rating to these than a non-precautionary one, but could still adopt and practice a risk-based approach overall.

In addition, in order to ensure trust and thus build support for their approach, regulators need to ensure that, whenever a decision is based on scientific advice or findings, they also set forth clearly what are the different values at stake, and if there is any actual or potential conflict between them. Acknowledging such conflicts is far more conducive to constructive engagement from all stakeholders with the advice given, whereas denying them by presenting the implicit values that form the advice's foundation as the only possible approach is creating strong negative reactions[455] (most scientific advice takes it as a given that safeguarding as many lives as possible is the main goal – but there are other values, like freedom or specific religious rules, for instance, that many citizens may see as deserving as much, or more, consideration – in other cases, advice incorporates an implicit "cost effectiveness" element, without discussing the alternatives, etc.).

In addition, it is important that "scientific advice" is understood not only as input from natural sciences into policy decisions involving technological and natural risks, but also as taking into account social sciences. This means first having social sciences give input into the policy advice on those policies aiming to address technological and natural risks, to ensure that issues related to behaviours, compliance etc. are adequately addressed, and that thus the presentation of policy options and their likely effects is realistic (see Wilsdon 2014). This is of major importance because, as a result, policies in social matters tend to be in many cases based far more on preconceptions and ideologies and far less on evidence. This also has serious implications for the legitimacy of all scientific advice and public support to evidence based policy: as long as it appears to be "cherry picked" and apply only to some issues, it is far more difficult to build broad-based consensus for it[456]. Considering findings from social sciences in regulatory matters is, precisely, what "smarter inspections" are about.

c. Applying risk-based approaches to inspections

Having attempted to summarize some of the main issues pertaining to the interactions of risk and regulation, and before we turn to examining in more details practical examples, it is time to consider the application of risk-based approaches to inspection from a general, part-theoretical and part-practical perspective. We will first consider the rationale for specifically basing *inspections* on risk (as distinct from "regulation" more broadly), then look at what are the main elements that appear to characterize "risk-based inspections" in the existing literature, and from there try to conclude on the theoretical basis for risk-based approaches in inspections and enforcement.

i. From risk regulation to risk-based inspections

[455] A policymaking process where the values and voices of stakeholders are not adequately represented will lose legitimacy – in contrast, procedural justice (irrespective of what the final decision is) will build legitimacy and thus acceptance of the policy decision in the end – see e.g. Maguire and Lind 2003.

[456] At the risk of being somewhat over-simplistic: much (but certainly not all) scientific advice on the risk of different products and technologies may end up showing that risks are acceptable and support broadly speaking "pro-business" policies. Broadly speaking "left-wing" groups tend to be skeptical of scientific advisory bodies as a result. Were scientific advice to also include social issues and social science, a quick look at the prevailing scientific evidence and consensus suggests that it may often result in supporting policies that are supported by these same groups that oppose the "pro-business" policies. By demonstrating that scientific advice and evidence is not "cherry picked" but used throughout all policy areas, it could contribute to broader support and acceptance, based on procedural justice effects (see above).

Risk-based inspections are not just a narrower field, a sub-section of a broader "risk and regulation" field – they are also one that has specific drivers, concerns and tools. While overall it is not only possible but legitimate to put them in the perspective of the broader risk regulation studies, it is also essential to understand this specificity.

BRDO's 2012 *Common Approach to Risk Assessment* outlines the different levels at which "risk assessment" (and, more broadly, risk-based approaches) can be applied in the regulatory sphere: "strategic risk", "priorities between national and local risk", "operational risk", "risk assessment of individual businesses" and "sanctioning according to risk" (pp. 3-4). In this outline, "strategic risk" corresponds to the overall strategy of the regulatory body, the key risks that it is its mandate to control. The setting of national or local priorities is in some ways UK-specific, given the importance of local regulation in the British system (though this articulation of different priorities can also be relevant to many other countries). "Operational risk" refers to the level where interventions are designed, the choice of regulatory instruments and tools made. The two last stages correspond to the classification of establishments according to risk (which is the basis for risk-based inspections, but can also be used for e.g. licensing), and to risk-proportionate enforcement.

We want to suggest here a slightly modified version of these different levels of risk-based regulation[457]: strategic risk assessment, operational risk assessment, risk-based targeting and risk-proportionate enforcement. The first deals with the policy-making level: what risks to regulate, and how. The second deal with the choice of implementation methods: what regulatory tools to use for which risks and situations. The third covers the targeting of inspections (and possibly of other regulatory tools). The last one deals with enforcement. While this classification is clearly based on that elaborated by BRDO, we think it introduces useful nuances and is more broadly applicable.

Mertens (2011) suggests a classification that focuses more on the risk assessment and management stages that take place within an inspectorate (p. 271). His classification has two broad levels. First a systemic one, which corresponds to the strategic inspection framework, defining priorities and programming. The output of the systemic stage is a classification of categories of risk level per type of establishment, and an action programme. Second, an operational level, which corresponds to the operational organization of inspections, involving information gathering and definition of specific focus. The output of this stage is an overview of results of prior inspections for each establishment, and an inspections plan.

There are several fundamental differences between what one could call the "macro" (strategy), "meso" (operational) and "micro" (targeting, enforcement) levels. While at the first level policy-makers operate at a rather high level of abstraction, and take decisions based on overall highly "aggregated" risk assessments, the "meso" level is already concerned with more concrete situations, and decisions that will translate into concrete differences for businesses – depending on which regulatory instruments are selected for different categories. The "micro" level, in turn, deals with individual cases (businesses, establishments), allocates them in one or the other category, and takes decisions based on findings on the ground. Thus, the first crucial difference is that one goes from abstract categories to individual cases.

The second essential difference is that the operational and "individual cases" levels all operate *within* the framework given by the strategic risk assessment. If the policy decision has been taken that a given category of risk will be addressed through regulation, then this is a given, which forms the environment within which lower level assessments and decisions will be made. The strategic level is where analysis such as Regulatory Impact Assessments can take place. At the operational level, the question is not anymore whether to regulate, but what instruments to use in order to best implement a regulatory decision already taken, with the available

[457] This relies on an internal World Bank Group paper, which was developed jointly with Wafa' Aranki and Lars Grava, both of the World Bank Group. They deserve equal credits for this.

resources, and taking into account what is known of the target groups, of compliance drivers, etc. Finally, at the individual cases level, the decision concerns first allocation of resources (given finite staff-hours, where will they be most useful) and how to respond to a given situation.

Whereas policy makers can, in practice, decide to regulate even if a realistic assessment would indicate that resources are insufficient, the regulation ill-designed, the goals unachievable – inspectorates cannot stretch resources beyond what they have. Thus, if they do *not* prioritize they will generally end up having to visit an unrealistically high number of premises (or check an unrealistically high number of products), meaning that each inspection will have to be very short. There are only three possibilities in the absence of risk-based planning: "blanket" coverage (every establishment/product is controlled), random inspections, or selection on a basis other than risk. In the first case, each inspection will have to be so short (except in rare cases of inspectorates dealing with a very small field) that it will be essentially useless – and, in fact, *within a given establishment* inspectors will be unable to control everything, hence there will be selectivity anyway (by default). In the second, there is formal "equality" (everyone has equal chances of being visited), but no uniformity in fact (some are visited, some not), and a clearly less-than-optimal resource allocation. In the third case, in the absence of a rational, somewhat objective instrument for selection, inspections end up being targeted based on convenience of inspectors, potential for flattering numbers (of fines, for instance), or rent-seeking. In other words, risk-based inspections are not an alternative to "non-selective" inspections, but to "selective by default" (see e.g. Blanc 2012, p. 31).

ii. Understanding what "risk-based inspections" entail

We have used so far a variety of related expressions to refer to our field of research, reflecting the diversity that is in use among both scholars and practitioners, and the different aspects that "risk-based approaches", broadly speaking, can take when applied to inspections. It is time for us to both specify more narrowly how we understand these different terms, and to consider what practices these refer to in the inspections field, based on the existing literature (both academic and originating from international organizations or state institutions).

"Risk-based inspections" are the broadest term: it refers to inspections approaches and practices that, generally speaking, are based on the notion of risk, and the idea that the regulatory response should be linked to the assessment of risk. "Risk-based planning" or "risk-targeted inspections" refer to the practice of linking the planning of inspection visits to the risk assessment of individual establishments (or, at least, of groups of establishments) – in one form or another, it is probably the most widespread form of "risk-based inspections", and also the meaning that most authors and practitioners are likely to associate immediately with the qualification "risk-based". "Risk-proportionate" inspections and/or enforcement refer to practices linking what is checked during the inspection visit, the importance given to different issues, as well as the way the inspection is followed up on (including enforcement decisions, if any) to the level of risk as assessed "on the ground".

Finally, there is a more comprehensive understanding of "smart", risk-based inspection practices that has not really been named adequately to date, and includes risk-based planning as well as risk-proportionality, but also goes beyond to incorporate a risk-differentiated approach in terms of selecting tools for compliance promotion (i.e. not only relying on inspection visits), and a far greater emphasis on information and guidance. This approach is grounded on a complex vision of compliance drivers, and seeks to make use of all of them at the same time. It corresponds to what the British now call "better regulatory delivery" (for which the Better Regulatory Delivery Office is responsible – but this is *even broader* since it includes other regulatory

instruments than inspections, e.g. licensing etc.). In specific discussions (when we try and "disaggregate" terms), we will refer to this as "smart inspections" – but in other cases, we will understand this to represent the "fullest expression" of "risk-based inspections". In other words, a *fully risk-based* approach to inspections will include *targeting* based on risk-assessment, *focus during visits* and *enforcement decisions* proportional to risk, and *compliance promotion approaches* which are differentiated based on risk (and on compliance drivers analysis). We will now take a slightly closer look to these different elements.

Risk-based targeting and planning

Before discussing the specifics of the criteria and tools used for risk-based targeting, a short preliminary discussion is required of a closely-related issue: the question of *reactive* versus *proactive* planning of inspections.

Reactive and *proactive* inspections

Inspection agencies can visit establishments either because they respond to a complaint or request (or a tip-off of some sort), or on the basis of their own planning, without any external trigger. Following a distinction introduced by Black (1970) for police work, Tilindyte (2012) refers to these two ways in which inspections can be initiated as "reactive" and "proactive", and this is a terminology used by a number of regulatory agencies themselves, at least in the EU. In other parts of the world, different labels may be used – in former Soviet countries, "proactive" inspections are known as "planned", and "reactive" ones as "unplanned". Whatever the words used, the distinction is widespread, and the vast majority of inspection agencies we have studied had a combination of both reactive and proactive work – but with very different proportions of each. In some rare cases, inspectorates even function nearly exclusively on the basis of complaints (reactive inspections).

Having reviewed the existing literature, as well as considered the issue from a theoretical perspective, Tilindyte (2012) comes to the provisional conclusion that complaints are more cost-effective, but that "only a small proportion of OSH violations are likely to come to the labour inspectorates' attention through private complaining". By contrast, "proactive policies (…) enable a more comprehensive, preventative and systematic approach to inspection" (pp. 42-43). Considering the specific experience of England and Wales, she concludes that inspectors mostly "do not view complaints as especially helpful" as "many of them are ill-informed" (p. 120). Inspectors in Germany reported problems linked to a "high number of complaints" coming from "disguised competitors" (p. 180). In other words, the *quality* of complaints-based information is frequently problematic in the OSH sphere.

If we consider other areas, the information basis for reactive inspections appears just as problematic, although in different directions. As shown by Bentata and Faure (2015), environmental complaints by private persons tend to be strongly biased towards "nuisances" rather than very significant pollution issues, and cannot form a sound basis for enforcement activity (and while their work shows NGOs picking up a significant amount of the serious cases in France, it cannot be assumed that this will be the case everywhere). In consumer issues, van Boom and Loos (2007) show that in the cases of repeated infringements with only limited loss for consumers ("trifle loss" problem), there is generally under-litigation (and, frequently, under-reporting). The propensity to complaint, in addition, is strongly linked to a number of social and cultural parameters. A recent OECD study of regulations in Lithuania (2015) shows that there is a real problem of excessive use of reactive inspections by the market surveillance inspectorate, and that the vast majority of complaints are trivial, or relate to issues that are not regulated by law (pp. 133-134).

"Proactivity" and "reactivity" are linked to the issue of risk-based targeting (or its absence), but in a somewhat complex way. In principle, complaints and other "tip-offs" *can* and *should* be integrated within a well thought-through risk-based targeting model. In practice, however, inspectorates that rely very strongly on complaints tend to have a very weak risk-orientation, if any. While Black (1970) focused on the bias in registering crime, there are a variety of major biases in *complaints*. These biases result from different cultures and perceptions, cost-benefits issues, social position (conditioning ease of access to "formal channels"), relationships with the objects of the complaint, etc. It can in no way be assumed that complaints will yield valuable information: many of them may be trivial, some are likely to be malevolent and dishonest, and if an agency simply follows up on each and every one of them, it will be stretched so thin that it may be unable to properly respond to the important ones.

As we will consider in more details in the third part of this research, inspectorates such as the Lithuanian market surveillance for which reactive work makes up more than half of all inspections tend not to be the "best practice" around. Systematic follow up of all complaints by an inspection tends to be frequent in post-Soviet or post-Communist systems, e.g. in Mongolia where more than 60% of the visits conducted by the General Agency for Specialized Inspections (GASI)[458] are "unplanned", i.e. complaints-based.

Rather, external sources of information can and should be incorporated into a risk-based analytical mode. As Tilindyte (2012) shows, this is the case for OSH in England and Wales, where the Health and Safety Executive handles complaints based on a series of factors, which allow to determine whether an investigation should take place (p. 119), and which include the potential or actual harm, past performance of the establishment, enforcement priorities, etc. A risk-based consideration of complaints can also take into account the existence of other (previous) complaints relating to the same establishment (which can be part of "past performance"), as well as the degree to which the complaint is substantiated. Conversely, in order to have up-to-date risk information on each establishment, an inspectorate needs to try and incorporate not only complaints, but other sources of information – coming from other inspectorates, the media, internet monitoring etc.

On balance, however, it appears clear that a risk-based approach to inspections means that an overwhelming majority of inspections would be proactive, and data-driven, rather than reactive and complaints-driven. It is worth noting that this relates to one of the key differences between regulatory inspections and police work (and more broadly crime-fighting work): most of the objects of inspections (establishments) are known, and the issue is to manage to estimate their risk level – whereas in criminal matters, identifying the culprits is precisely the main problem (and, in "victimless crime", identifying the crime itself). This is not to say that detection problems are not important (cf. Baldwin and Black 2008), but (at least for the most relevant inspection functions in terms of numbers), the universe of establishments is known, and the planning task is to determine where to go in priority. The primary objective is prevention, not response (even though response also matters). By contrast, even though police work aims *overall* at preventing (reducing, containing) crime, its operational focus is to a large extent based on *response* (even though of course there is a large amount of preventive action, e.g. patrolling). For these reasons, the significance of the *reactive* work as identified by Black (1970) for police work is far lower for inspections[459]. Bardach and Kagan (1982) make this very same point that "enforcement of protective regulation by inspectors is different" from typically law enforcement as

[458] The GASI gathers most inspection functions, except fire safety and revenue (tax, customs). Based on internal (unpublished) GASI data for 2013 and 2014.

[459] We posit here a strong difference between "regulatory inspections and enforcement" and "police work/criminal law enforcement". This difference is far from always being obvious, there are many "grey areas" and complex interrelationships, but on balance we think that the difference in fundamental focus is meaningful. It would have to be further investigated and discussed in future research. It is worth noting that we are far from the first to make this point, and to note that criminal law approaches are not necessarily the most effective or efficient for regulatory issues (see for another perspective on this Simpson 2002, investigating what she calls the "punitive model of corporate crime control" (p. 10), and concluding to its inadaptation to business regulation issues).

"inspectors sometimes respond to complaints, but they usually come on their own initiative to enterprises that have not been accused of any wrongdoing. They search for *ongoing* violations, things that might go wrong in the future" (p. 31). The question then is how best to select these places to proactively visit.

Targeting and planning in practice – the data issue

Selecting enterprises to be (proactively) inspected based on their risk profile is so essential to risk-based approaches that the two are often identified, i.e. that many lose sight of the fact that a proper "risk-based approach" includes *more* than targeting. We have already outlined above that the foundation of risk classification for inspections is to combine the *likelihood of harm* with its *potential severity and magnitude.* Doing so in an effective way requires disaggregating the processes that may lead to harm, in order to understand what are the causes of the harms that the inspectorate seeks to prevent, and to ensure focus on the right issues and establishments (cf. Mertens 2011 pp. 272-273).

There are different ways to structure the classification that is to form the instrument for planning. One approach (BRDO 2012, World Bank Group 2013 a) is to form a matrix with two axes – one corresponding to the likelihood of harm (including the likelihood of non-compliance, but not limited to it – World Bank Group version) or the likelihood of violation (BRDO version), and the second to the potential severity and magnitude. In such an approach, intrinsic risk and management risk are somewhat aggregated in the way they are presented (even though, analytically, they are to be handled separately). Another is the approach presented by Mertens (2011, pp. 273-274) where the risk classification is done purely on the basis of intrinsic risk, and *then* a level of inspection priority is determined by crossing the resulting risk level with the compliance history or expectation (management risk).

In any case, a fully-fledged risk-based targeting is to take into account a set of risk components that includes (a) intrinsic risk of the activity (hazardousness), (b) scope/size of the activity (number of people who could be affected, or other relevant indicator), (c) additional relevant vulnerability factors (e.g. types of populations affected, location etc.), (d) likelihood of harm. This last element can be itself split between intrinsic likelihood (which can be combined into "intrinsic risk", or not – in which case intrinsic likelihood and intrinsic severity are handled separately) and management-related likelihood, or "compliance risk". The relative weight that is given to each of these factors can vary (even though most "matrix" models suggest that severity and likelihood should overall be given equal consideration, precise methodologies are diverse). The ways in which these are rated, graded, measured etc. also varies considerably, with some agencies having far more sophisticated and "data-driven" models, some far more "qualitative" approaches (see Baldwin and Black 2010). The use of "qualitative" indicators does not mean the rating systems are necessarily simple – the Food Standards Agency in England and Wales has indicators that are mostly not data-driven, but a rating system that incorporates a number of dimensions and a sophisticated set of check-lists (cf. Blanc 2012 p. 33).

Once a classification has been created, as well as a grading/rating tool to assign a risk rating or category to each establishment (or product, or more generally "inspection object"), *targeting* involves assigning a category or rating to each *concrete* object, and to decide on an actual *plan* of inspections. These are two conceptually separate processes (whichever way they are actually conducted in practice).

If we look first at the question of *planning*, it involves matching resources to the needs, establishing "typical frequencies" for different risk categories, and also adding (or not) an element of "random selection". Here the most practically logical approach would involve deciding first only on an optimal frequency of visits for the highest risk category, then adjust it downwards if existing resources do not allow to implement it, and only then look at *factually possible* frequencies for lower categories, given existing staff and average duration of inspections. There are, however, many cases of frequencies assigned for all categories based on more-or-less

arbitrary estimations of what is "adequate", which then may or may not be feasible given available resources. In any case, the guiding principle of a risk-based approach is that high risk establishments should be visited far more regularly and frequently, that low-risk ones may not even warrant regular visits at all, and that the classification should be done in a way that results in only a minority of businesses being at the "peak" of the "risk pyramid" (World Bank Group 2013 a). In order to keep a "reality check" of whether the classification and ratings are adequate, and to avoid creating incentives for non-compliance for low-risk businesses, it is often accepted that keeping some level of (rare) randomly selected inspection cases for the low-risk category is a valid approach (see Baldwin and Black 2010, Sparrow 2008 *et al.*).

While the classification and the "indicative frequencies" for each category provide the tools for the actual planning, replacing abstract categories with actual "targets", establishments to be visited, requires *data* as a foundation – at a minimum, a list (database) of all establishments under supervision, with at least some fundamental information on the most important parameters that allow to determine the risk level. In some cases, the database can be very sophisticated, and be paired with an automated case selection system (which also takes care of matching frequency of visits with available resources, etc.) – in most cases, the systems are less sophisticated and require a significant human input. In any case data is, in a number of jurisdictions and agencies, the weakest link. There are, however, frequent misconceptions around this, so it is important to distinguish what is absolutely necessary from what is "good to have", and to understand what is the real level of operational challenges and resources involved.

A common assumption is that putting in place effective data systems for risk-based targeting and management of inspections would be very costly, and that moving to a risk-based approach is thus a major investment for an inspectorate – which, in turn, can be a reason to settle for avowedly inferior approaches to inspections. Such an assumption underlies for instance Tilindyte's statement that "proactive monitoring" has "generally high costs", "especially if it is to be based on a comprehensive risk assessment" (2012, p. 42). Baldwin (2007) expresses similar concerns, with more specifics: "a further tension (…) may arise out of the Government's desires (a) to reduce quite significantly the burdens of supplying information (…) and (b) to ensure that regulators target their enforcement activities more precisely (…) The problems are, first, that the targeting of enforcement demands that inspections and other actions are based on intelligence and, secondly, that if the obligations of businesses to supply information to regulators are reduced, it is increasingly difficult for regulators to engage in targeting without generating intelligence independently. Such independent generation of data may, of course, prove hugely expensive for regulators – indeed far more expensive for them than for the businesses that they are controlling" (p. 40). There are many points here, which all deserve to be properly addressed.

First, *in theory*, it may be true that building an information database on objects under supervision and a risk-based targeting system *from scratch* may be expensive. Similarly, regularly gathering information "in a vacuum", i.e. launching extensive investigations, would certainly be costly for an inspectorate. In fact, *any form* of information gathering has costs (even processing data submissions by businesses), and there is no doubt that making a planning system *more data-driven* will increase somewhat data-related costs. Finally, *assuming the information submitted by businesses is adequate*, it is clearly cheaper for the regulators to push the information collection burden on them. All these points, however, rely on assumptions that are fundamentally at odds with reality, at least as observed in most cases.

The first inaccurate assumption is that such information database would have to be built from zero – in fact, most inspectorates around the world have been operating for years or decades and, at least in OECD countries, the vast majority already have databases of objects under supervision, even if these may be managed through sometimes outdated software, or be partially incomplete etc. Furthermore, gathering data on establishments under supervision is something that *naturally occurs anyway* as part of each inspection. Given that, in fact, the inspection coverage tends to be far higher than suggested by studies focusing on only one agency (sometimes

"marginal" in terms of volume of activity), inspectorates gather each year a considerable volume of data simply as part of their normal control activities. The problem is that this data is often not managed properly, i.e. not entered into systems that would make it useful for further analysis and planning. Another difficulty is that in many contexts inspectorates do not share information among themselves, which reduces the number of establishments they can cover in a given year (cf. Blanc 2012 pp. 21-25 and 77-80). In other words, the already are considerable "sunk costs" whereby inspectorates have collected or are regularly collecting information, through their main activity i.e. inspections – the problem is how to best make use of this existing data.

A second highly problematic assumption is that whatever information is filed by businesses will be accurate (if not fully, then at least mostly). This is, based on our experience, unlikely to be always true, and in fact unlikely to be the case *precisely on some of the businesses where information is most needed*. Indeed, if we come back to our compliance models, and the proposed typology of different profiles, precisely the establishments which are the least likely to comply are also the least likely to submit truthful information, as they will correctly understand that this information may be used for risk-based profiling[460]. As for those who are inclined to voluntary compliance, burdensome information collection is likely to create resistance and to overall lead to a *decrease* in compliance. Thus, it is unlikely that relying strongly on information submitted by businesses themselves is ever a very good idea. It is, in some cases, relevant and necessary – but it should remain simple, and certainly not be the sole (or even the main) source of data[461]. This is not to say, again, that business-reported data cannot be useful – but that in any case it never could be sufficient. This is particularly obvious if one thinks of the case of "fly by night" businesses, i.e. those who try to stay invisible and operate partly or fully illegally, and without control. Both Sparrow (2008) and Baldwin and Black (2008, 2010) discuss in some depth these cases. Clearly, detection of such businesses will not be improved by relying on reporting obligations[462]. Rather, inspectorates need to rely on a combination of tools to "spot" businesses operating "under the radar": tip-offs and complaints, "physical" monitoring (verifying whether visibly operating premises are listed in the database), online monitoring (looking for signs of activity, e.g. websites, advertisements or social media comments, and checking whether the business is listed), and information sharing between regulators (if one of them detects an unregistered business, all of them should be notified). This shows how much active data collection is, in any case, a condition of effective supervision, with or without risk-based approach.

More effective sharing of information between different state bodies (and in particular between those which have a regulatory and/or supervisory function) is indeed an essential element of "smarter regulation", if by this we understand a way of regulating that would be *both* more efficient and more effective. This is particularly true when it comes specifically to risk-based inspections – information sharing is key to improving data on establishments/products under supervision, and making sure risk information is comprehensive and up-to-date. It is important to remind that, again, this is not *only* linked to the introduction of risk-based inspections. A number of governments have put in place, or are trying to introduce, policies or tools to avoid duplicating information requests, and ensuring that information collected once is shared across all of the public administration[463]. Some examples include the abolition of the use of certificates in the relations

[460] See a detailed account of such problems in Bardach and Kagan 1982, pp. 90-91 – they write, in summary: "Documentation, by its very nature, is a declaration of innocence, and most of it is received by officials who ignore it almost entirely" (p. 90).

[461] Enterprise-submitted information is primarily useful to simply notify existence of an establishment – through business registration (for all establishments) or for specific activities (e.g. EU-wide notification of food business operations). These are also relatively "risky" for the business to evade, at least if the activity is easy to detect. Another case where information is more likely to be truthful is high risk, large scale businesses where inspections are frequent. In fact, for such businesses, relationships with regulators tend to be "ongoing", and this is not really the target group for reducing reporting requirements.

[462] One could argue that punitive sanctions for non-reporting could strengthen deterrence, but for all the reasons exposed in the compliance section, this is not very likely to work for most cases.

[463] As much as privacy legislation allows. In some cases, privacy concerns and/or applicable laws have been making information sharing more difficult, but this is a concern that is stronger in the case of citizen rather than business information, and in any case goes beyond

between private persons and the public administration in Italy[464], or the Netherlands' *Stelsel van Basisregistraties* ("System of Basic Registrations", in other words a system of unified registries[465]). More recently, France has embarked in a similar direction with the programme "*Dîtes le nous une fois*", which aims at avoiding duplicate requests and submissions of similar information[466]. At the EU level, important instruments have long been helping with exchange of information on emerging hazards in food and non-food product markets (RASFF for food and RAPEX for non-food[467]).

Specifically in the inspections field, information sharing can be done in a number of ways. At the local level, the fact that most inspection fields are under a single department under local authorities in the UK means that there is a good amount of information sharing going on between them, and there is ongoing work to develop information systems that will make this sharing more systematic and easier – and also ensure that sharing happens *between different regions*[468]. In the Netherlands, two systems have been developed to allow for more effective sharing of information between inspectorates (*Inspectie View*[469]) – and to allow inspectorates to access a trove of data on the business, avoiding duplicate submissions, specific queries etc. (*Ondernemingsdossier* – "Enterprise File"[470]). In Italy, a somewhat similar system has been created, first at the regional level (in Emilia Romagna since 2011) – with an extension to the national level now decided upon – the *Registro unico dei controlli* ("Unified Registry of Inspections") for the agricultural (and agricultural processing) sector, which allows inspectors of all relevant agencies to see records of all inspections, even by other agencies[471]. Clearly, much is happening in this direction – however, the existence of many legacy systems and institutional barriers mean that integration is done *ex post*, in a relatively uneasy way, and without automation (it all relies on inspectors actually using the system to make queries). The *Inspectieloket* portal even suggests that the decision to have different *Inspectie View* for different domains was done to avoid

---

the scope of our research at this stage. The very real trade-off between privacy and burden is often poorly perceived, and this is clearly an area where more efforts should go both in terms of research and of policy discussions.

[464] Requesting certificates is in fact *prohibited* and would be a violation for civil servants – as per Law n. 183 of 12 Nov. 2011. See explanation on the website of the Office for Administrative Simplification: http://www.funzionepubblica.gov.it/lazione-del-ministro/decertificazione---direttiva-n-142011/la-direttiva-del-ministro-per-la-pubblica-amministrazione-e-la-semplificazione.aspx.

[465] See detailed presentation of the system here: http://www.digitaleoverheid.nl/onderwerpen/stelselinformatiepunt/stelsel-van-basisregistraties.

[466] A principle that is inspired e.g. by previous experiences in the Netherlands. On the French programme, see the following website: http://www.modernisation.gouv.fr/les-services-publics-se-simplifient-et-innovent/par-des-simplifications-pour-les-entreprises/dites-le-nous-une-fois-un-programme-pour-simplifier-la-vie-des-entreprises.

[467] See on the European Commission website: http://ec.europa.eu/food/safety/rasff/index_en.htm on RASFF, and http://ec.europa.eu/consumers/consumers_safety/safety_products/rapex/index_en.htm on RAPEX. See also the first chapter of this research on the creation and development of RASFF.

[468] Which, in the past (and even currently), has been a significant problem at least in some areas. See e.g. Ogus, Faure and Philipsen (2006) p. 40, which underlines the problem with various risk-assessment models (something the 2012 BRDO *Common Framework* was precisely created to address).

[469] The system has gradually developed over several years and was created to allow any inspector to access data from other inspectorates on a given establishment/object – in particular prior inspection records. There is a general level (*Inspecte View Bedrijven* – "Inspection View for Companies") which can be used for planning and aggregates inspections and results from Social Affairs, Environment and Transport, Food and Non-Food Products Inspectorates. There are then several "specialized *Inspectie View*" with a deeper level of information sharing (greater wealth of information, e.g. on permits etc.) – for inland transport and environment (for now). For more information, see *Inspectieloket* portal: http://www.inspectieloket.nl/organisatie/index/ - and detailed files at the project webpage: http://www.informatieuitwisselingmilieu.nl/publicaties.php?id=11.

[470] The "Company File" allows to access all the information the company decides to make available – it is being rolled out gradually, by sub-sector of the economy, as it is run by businesses, not by the public administration. More information is available at: http://www.ondernemingsdossier.nl/. The "Company File" can be seen as an attempt to not only avoid duplication of reporting requirements, but also to access *more* information from businesses and thus make overall planning and targeting more effective (see Baldwin and Black 2008 p. 31 on the importance of mobilizing the private sector in gathering information).

[471] For the Emilia Romagna experience, see on the Region's portal: http://agrea.regione.emilia-romagna.it/servizi/accesso-agli-applicativi-1/registro-unico-dei-controlli-ruc. For the decision to expand it nationally, see on the Ministry of Agriculture's portal: https://www.politicheagricole.it/flex/cm/pages/ServeBLOB.php/L/IT/IDPagina/8631.

"unnecessary and excessive complexity and size of data". This appears, if one considers the "state of the art", to be more of a fig leaf than a genuine problem.

In fact, several jurisdictions have gone much further and created fully integrated databases for most inspection types (generally excluding fiscal ones), linked to a management system that directly uses the data and risk management guidelines to produce an inspection plan (and in some cases even assign cases directly to inspectors) – not to mention other features for inspections results recording, data analysis etc. These cases and best practices, which are mostly to be found in reports prepared by international organizations such as the World Bank Group (2014 b) and the OECD (cf. Blanc 2012 pp. 77-80), mostly (though not exclusively) come from emerging markets (broadly defined). With a lower presence of legacy systems, relatively lower institutional resistance, and the rapid technological progress (which lowers costs from year to year), it has been possible to set up systems that are far more advanced and effective. In fact, what appeared particularly difficult and costly a few years ago is now far more feasible (e.g. having a fully integrated database across most inspectorates) – but it requires significant decisions (political and technical), and good management.

In short, the important conclusion is that even really advanced and integrated systems are increasingly "feasible", and with certainty simpler systems of data collection and management are fully possible to implement even with relatively constrained budgets. Of course, any data collection and management system will have costs, and implementing analysis-driven planning will have costs relative to "rule of thumb" targeting – but these costs are far from being as considerable as suggested by several authors (which maybe relied too much on the testimony of regulators themselves, who may have their own motives for being reluctant), many data collection activities are anyway necessary (and it is just about using this data more efficiently), and there are considerable costs (in effectiveness) in the *status quo*. As we have noted above, such efforts to achieve more consolidation and sharing of data can be challenged based on privacy concerns (and privacy and data protection legislation, in some cases). There are very different perspectives on what is the appropriate level of privacy and data protection in different countries, and it is obvious that implementing such new systems would be more difficult e.g. in Germany than in the UK, from this perspective. Because the information at issue is *corporate* rather than *personal*, and because of the overwhelming case to be made from an efficiency and effectiveness perspective, we do not think such concerns should stand in the way of data sharing in the field of regulatory inspections of economic activities (as distinct from other areas where data sharing may be considered, and which are not the object of our research). Indeed, in countries where efforts at data consolidation and integration have been made (the UK, the Netherlands with *Inspectie View* and the "Company File", Italy with the *Registro Unico Controlli*, etc.), the parties directly affected (the businesses) have been in favour of the change, and have not generally voiced concerns. It remains that it may be different in other contexts (e.g. regarding the publicity of inspection findings such as is the case for food hygiene ratings), and that the issue is not uncontroversial. Technically feasible does not mean legally feasible, and does not mean desirable either (though, from an instrumental perspective, it clearly is).

Risk-based inspections "on the ground" – risk-proportionate enforcement

As we indicated above, targeting and planning are but the first element of risk-based inspections, and the way inspections are actually conducted "on the ground", as well as the way inspectors and their management *follow up* on them, are just as essential if one is to have an approach that is really founded on risk. While there have been very important scholarly works focusing on how inspectors take decisions and interact with regulated entities, there has been rather less on *what they check* (and what skills, experience and culture influence it). There has been considerably less work on inspections and enforcement practices *specifically focusing* on risk-based approaches, how inspectors understand them, and how they are translated into practice. Hawkins's very important work on enforcement practices in Britain's Health and Safety Executive (2002) considers in great detail and depth the practices of inspectors, the framework which influences their

decisions, the ways in which the agency's management attempts to shape them. It does this, however, without a specific focus on the question of *risk*, but rather at the notion of *discretion* and with a very open investigation of all the drivers that may be at play – and with a specific focus on the enforcement rather than on the inspection phase[472]. Overall, a relatively "diffuse" notion of risk permeates both Hawkins's work and the culture and framework in which he sees inspectors as operating – but not (and this was not his research's purpose) a "picture" of what "risk-based inspections practices" may look like. Baldwin and Black (2010) seek to define "really responsive risk-based regulation", but focus more on the intermediate, operations management level, than on the inspecting stage. Similarly, Sparrow (2008) considers more problems identification and "harm control/reduction projects" than the work of control at the "end phase". May and Winter (2012) consider the relative effectiveness of "enforcement styles", but not whether risk considerations may play a role in it, and without really looking at the inspection phase. The same could be said in general of much of the work on enforcement, including Ayres and Braithwaite (1992) and others: the main interest is the effectiveness of different interaction and enforcement strategies, not what inspectors actually check when they conduct an inspection visit.

Among major earlier works in our field, the closest to our question may be Bardach and Kagan (1982), since their consideration of "regulatory unreasonableness" to some extent looks at what inspectors check (at least through the prism of "what they then decide to enforce"), and does it with a prism that is closely related to "risk-based approaches". Their definition of "unreasonableness" can be read as, in a way, the opposite of "risk proportionality": "a regulatory requirement is unreasonable if compliance would not yield the intended benefits (…) Further, a regulatory requirement is unreasonable if compliance would entail costs that clearly exceed the resulting social benefits (…) Finally, unreasonableness means cost-ineffectiveness" (p. 6). One of the book's first examples illustrates how "unreasonableness" could also be less effective *in absolute terms*, i.e. distract attention and resources from more important problems, when a business operator (aluminium smelter) says of the "worst case scenario" which OSHA uses to justify its requirement: "Of course it *could* happen. Almost anything *could* happen. Never mind that it's more likely that an earthquake could happen. (…) This is a total misapplication of resources. I could use that money for real risk reduction in plenty of other places" (pp. 4-5). The same interviewee in fact refers clearly to the question of risk assessment: "Never mind that in the 15 years the plant has been operating nothing like that happened, or even *any incidents that suggest it might happen*" (p. 4 – emphasis ours).

While Bardach and Kagan use the word "risk" only rarely, and do not use the "risk-based" concept (which was yet to emerge at the time), the portrait they make of the "good inspector" encompasses many of the fundamental aspects of risk-based inspections "on the ground". First, they present precisely the problem that is one of the key justifications for a risk-based approach to select *what to inspect*: "the inspector who walks through a factory and faithfully enforces each regulation may not detect or do anything about more serious sources of risk that happen to lie outside the rulebook; at the same time, he alienates the regulated enterprise and encourages noncooperative attitudes" (p. 123). Indeed, at the core of risk-based inspection work on inspected premises is the idea of effective investigation, looking for the key risks, which requires to know how to prioritize, what to look for, and how to stimulate cooperation in order to get insider information (or, barring this, to detect dissimulation, and act accordingly). Bardach and Kagan introduce their vision of the "good inspector" by analogy with the "good cop", whose goal is to "reduce serious crime, particularly crimes of threat and violence" (p. 125). Translated into the regulatory field, this corresponds to a strong focus on risk, on "harm reduction". In order to achieve this, the police and regulators both need *cooperation* – "good community

[472] We will return to the findings of Hawkins's work in the third part of this research, looking at current HSE practice and considering whether there has been any evolution compared to the period his work considers.

relations is an essential element of effective law enforcement" because "citizens must be willing to inform the police of serious law violations" (*ibid.*).

From these premises emerge the vision that a good inspector "must have sufficient knowledge and understanding" but also at the same time "certain personality traits and communications skills". S/he must have "the capacity to empathize with those subject to the law and to understand their concerns, problems and motivations" (p. 127). These "communications skills" and understanding of the establishments s/he regulates should enable to (as much as possible) gain "compliance without stimulating legal contestation" (p. 128). This requires a "critical ingredient": "the capacity to be reasonable, to distinguish serious from nonserious violations, and to invest effort in the former" – which, in turn, requires "technical competence" (including understanding "the technical and economic problems of compliance", so as to be able to "evaluate the businessman's excuses or complaints" – *ibid.*). The inspector must have "tough-mindedness to probe", "be willing and able to exercise authority", and be "patient and persistent in the face of resistance" (pp. 129-130). S/he must be ready to offer "forbearance to elicit compliance" (p. 136), being lenient on minor issues to achieve progress on more important ones. Gaining cooperation may also involve supplying information: "drawing on its cumulative experience with a variety of firms", the inspectorate "can provide information about risks and abatement techniques", and inspector can advise "about significant hazards that have escaped the attention of company officials" (p. 143). The advice will be particularly well received if it "enables to make reforms more cheaply, and with less disruption of routine" (p. 144). The key, in other words, is to have inspectors that are able to spot and help solve *problems* rather than focusing on *violations* (p. 79-80).

The problem is then how to *enable* such inspectors to arise, and to work? First, of course, this way of working should not be forbidden: "good inspection can flourish only in an organizational and political environment that cultivates it, or at least permits it" (p. 151). Further than this, there should be tools to help inspectors do their work, which involves making "intuitive judgments about the motivations and capabilities they deal with" (p. 71), and developing a "specialized vision, more sensitive to possible risks and deceptions than the average person's" (p. 82). Risk-based approaches have been developed *precisely* with the intent to enable inspectors to be more along the lines of the "good inspector" defined by Bardach and Kagan, to help them have more effective tools for detection, but also better skills and approaches both for investigation and to stimulate cooperation and compliance. We will give here just a few examples of what can be done to make such "good inspectors" better equipped, and more numerous – considering the examples of the Health and Safety Executive (HSE) in Britain (*Enforcement Policy Statement* and *Enforcement Management Model*), the UK Better Regulation Delivery Office (BRDO) *Common Approach to Competency for Regulators*, and Lithuania's experience with risk-based check-lists.

Through its *Enforcement Policy Statement,* the HSE sets out the goals of inspections and enforcement activities, and their key principles – with risk as an essential foundation. First, the goals: "The ultimate purpose of the enforcing authorities is to ensure that dutyholders manage and control risks effectively, thus preventing harm" and "The purpose of enforcement is to: - ensure that dutyholders take action to deal immediately with serious risks; - promote and achieve sustained compliance with the law; - ensure that dutyholders who breach health and safety requirements, and directors or managers who fail in their responsibilities, may be held to account" (p. 2). Addressing risks is thus the most important, ultimate purpose. Then, the *Statement* lists the tools to these aims: "The enforcing authorities have a range of tools at their disposal in seeking to secure compliance with the law and to ensure a proportionate response to criminal offences. Inspectors may offer dutyholders information, and advice, both face to face and in writing. This may include warning a dutyholder that in the opinion of the inspector, they are failing to comply with the law. Where appropriate, inspectors may also serve improvement and prohibition notices, withdraw approvals, (…), and they may prosecute" (*ibid.*). Proportionality is immediately put forward. The *Statement* goes on to define the *principles* on which inspectors (and the whole organization) should base their actions (and their choice of tools): "HSE believes in

firm but fair enforcement of health and safety law. This should be informed by the principles of proportionality in applying the law and securing compliance; consistency of approach; targeting of enforcement action; transparency about how the regulator operates and what those regulated may expect; and accountability for the regulator's actions" (p. 3). The principles are then defined in further detail: "Proportionality means relating enforcement action to the risks" and "In practice, applying the principle of proportionality means that enforcing authorities should take particular account of how far the dutyholder has fallen short of what the law requires and the extent of the risks to people arising from the breach" (p. 4)[473]. Targeting, while it relates primarily to planning (see previous section), also has implications for how inspections and enforcement are conducted in practice: "Targeting means (…) that action is focused on the dutyholders who are responsible for the risk and who are best placed to control it – whether employers, manufacturers, suppliers, or others". In order to address the problem of excessive discretion and lack of equal treatment, HSE has a principle of *consistency*: "Consistency of approach does not mean uniformity. It means taking a similar approach in similar circumstances to achieve similar ends" (p. 5). Finally, "Transparency means helping dutyholders to understand what is expected of them and what they should expect from the enforcing authorities" (*ibid.*).

Such statements may be quite difficult to put into practice and, in fact, Hawkins (2002) suggested that, while official enforcement policy was *one of the elements* forming the framework for enforcement decision-making, they were but one of many, and in practice they left much to interpretation by inspectors (and their managers). In the meantime, the HSE developed a highly detailed, specific and practice-oriented tool to implement its enforcement policy: the *Enforcement Management Model (EMM).* The EMM's purpose is to "promote enforcement consistency by confirming the parameters, and the relationships between the many variables, in the enforcement decision-making process", to "promote proportionality and targeting by confirming the risk-based criteria against which decisions are made" and to "be a framework for making enforcement decisions transparent, and for ensuring that those who make decisions are accountable for them" (p. 5). While it does not replace or limit inspectors' discretion, it aims to guide it (in particular for less experienced inspectors). The EMM includes a number of "decision trees", rating tables and matrices helping inspectors to make decisions based on risk. We will quote here only some of the most important elements. As a first step during inspections, "inspectors collect information about hazards and control measures. From this, they make judgements about the health and safety risks associated with the activity under consideration. Inspectors should prioritise specific hazards and consider common root/underlying causes to ensure they deal immediately with serious risks. They should consider how best to achieve sustained compliance with the law" (p. 8). Then, inspectors should *assess* risk: they "should always deal first with matters that give rise to risk of serious personal injury. They have the power to either prohibit the work activity, or seize and make safe the article or substances that are creating the risk. Sometimes they will do both. When considering the immediacy of risk, inspectors should

[473] Proportionality is also a guiding principle in more targeted documents, eg. the Health and Safety Executive's *Enforcement policy in respect to iron gas mains* (2005). The context of the adoption of this enforcement policy was public risk concern: "In September 2001 HSE published its enforcement policy for the replacement of iron gas mains for the period 2002 - 2007. This followed a high level of societal concern about the potential consequences of gas mains failure. At that time records showed there were about 91 000 km of iron mains within 30m of property ('at risk') which may be a risk to people. (…) Given the uncertainty about this issue, HSE undertook to review the policy before the end of the first five years so that an agreed programme could be confirmed for the following period. The HSE's conclusion was that it was unrealistic to replace all iron gas mains in a short timeframe, but that at the same time "there is currently no feasible alternative to maintaining the network other than to decommission it and replace it with a more suitable material, usually polyethylene. This is the basis of HSE's enforcement policy, which requires iron gas mains within 30m of property to be decommissioned and replaced at the latest by March 2032 ". Basically, the enforcement policy offers gas network operators the option of developing a replacement programme and, if HSE approves it (for which it must be ambitious enough), they will have serious benefits in terms of enforcement: "if pipeline operators have an approved programme, they have a defence from prosecution if they are complying with it and a failure occurred on a pipe which was not yet due for replacement under the programme. However, the defence would not apply if the operator had knowledge which would indicate that the particular pipe was likely to fail". The solution adopted does not remove the legal obligation to overall replace all these pipes, but accepts that there must be a timeframe to do so, and offers defence from prosecution to firms that work in good faith on addressing the issue. See the policy on the HSE website available at: http://www.hse.gov.uk/gas/supply/mainsreplacement/irongasmain.htm

use the principles of 'risk gap analysis'" (p. 9). This "gap analysis" is then explained: once inspectors have determined the "actual risk (where the dutyholder is)", they should "compare this to the risk accepted by the law or guidance and decide the benchmark risk(the level of risk remaining once the actions required of the dutyholder by the relevant standards, enforceable by law, are met). The difference between where the dutyholder is and where they should be is the risk gap" (p. 12). The risk gap is then combined with the "authority of the standard" (level of clarity, specificity, strength of the rule) in order to give an "initial enforcement expectation" (p. 24). Then, the inspector should consider "dutyholder factors" and "strategic factors" (p. 31), "the factors specific to a particular case which may vary the initial enforcement expectation". "Dutyholder factors" include the compliance history, prior enforcement (or lack thereof), whether the violations were caused deliberately to seek gain, what are the general conditions in the establishment, behaviour of the operator ("responsive" perspective) etc. (pp. 31-34). "Strategic factors" are considerably more vague, essentially meaning that inspectors should check whether "the proposed action will produce a net benefit to the wider community in terms of reducing risks, targeting public resources on the most serious risks and the costs of pursuing a particular course of action" (p. 40). For instance, public expectations of a "tough response" may lead to a more severe action, but socio-economic impacts may also suggest in some cases a less severe one.

The EMM is a significant step (and, to our knowledge at least, unique – at least in its specificity) in making inspections and enforcement simultaneously more risk-based, more responsive, and more consistent. To put such tools to good use, however, competent inspectors are needed. In fact, the more flexibility is introduced, the more discretion is needed, the finer the assessment of risk required – the more competent inspectors are indispensable[474]. This notion of "competency", however, includes more than only *technical* skills (relating to food safety, occupational safety and health, environmental protection etc.), but should also encompass skills relating to risk assessment, investigation, relations with business operators and their staff, compliance promotion etc. In the UK, a model has been developed in recent years, building on work done within HSE. This effort involves a number of regulators and professional associations of regulatory staff is led by BRDO and has produced a *Competency Approach.* This is based on a set of "core skills" that are complemented by "technical skills" (rather than seeing core skills as "soft skills" they are put first). Among the core skills are: "assessing risks", "planning", "promoting compliance", "advising and influencing", "interventions", "enforcing legislation", "work with business", "work with partners", "using knowledge", "personal development" and "IT Literacy and Numeracy". The importance of skills relating to risk-based approaches broadly understood, including the choice of interventions, cooperation and persuasion, risk assessment etc. is particularly clear. The approach is fundamentally turned towards practice, and is thus not articulated in any lengthy document (only short summaries exist), but rather is supported by two web portals. The first is used for self-assessment (Regulators Development Needs Assessment – RDNA[475]) and the second for information and training (Guidance for Regulators – Information Point – GRIP[476]).

We have presented examples of clearly sophisticated approach, from a country where arguably risk-based approaches to regulation and inspections are the most established. It is important to consider whether such approaches are also applicable and realistic for countries where leaving discretion to inspectors can be associated with greater fears of abuse, where competency and professionalism are somewhat lower, and where the legal and regulatory culture generally is different. While we will present more examples and discuss this issue in greater depth in the third part of this research, looking at one short example will help complete this section. In Lithuania, since 2010, an ambitious programme of inspections reform has been underway,

[474] Badarch and Kagan (1982) showed that, conversely, insufficiently competent inspectors tended to "go by the book" (pp. 128-129) and be both more "unreasonable" and less effective at managing risks.
[475] http://rdna-tool.lbro.org.uk/
[476] http://www.regulatorsdevelopment.info/grip/

openly modelled on the UK experience (and drawing more broadly on international experience and lessons – see OECD 2015 b). As part of this reform, the Government has promoted the development and use of check-lists by inspectorates, in particular for inspections of SMEs. This was first requested by a Government decree, and is now also part of the amended Law on Public Administration. In addition, the Ministries of Economy and Justice adopted guidelines for inspectorates on how to develop such check-lists, emphasizing the need to design them based on risks, and not by compiling all applicable legislation. The aim is to have problem-oriented check-lists, that guide inspectors to look at the most essential issues, where the most risk can arise, and take them away from "paperwork-focused" inspections[477]. Interestingly, check-lists for inspectors are not seen positively in more "advanced" inspectorates (e.g. in the UK, or in some agencies in the Netherlands) precisely because they are seen as excessively limiting discretion, leading to a "tick box" approach, insufficiently promoting professionalism. In Bardach and Kagan's (1982) account, check-lists were in fact a tool that had been introduced as part of the more rigid, more "protective" regulatory approach that emerged in the 1970s (pp. 74-75), and check-lists were generally examples of "zero discretion" practices, leading to "regulatory unreasonableness". Here, two factors are essential to consider: context, and contents of the check-lists. Context, first: a system where risk-based approaches run against deeply engrained practices of inspectors, and where resources are not necessarily available for in-depth retraining or to attract new and more qualified staff. In such a case, well designed check-lists, while not "optimal", can represent a major improvement by pushing inspectors to a somewhat simplified but still adequate risk-based practice. Contents, second: poorly designed check-lists will indeed end up with hundreds of items, a laundry list consisting of many paperwork requirements and lending itself to "by the book" enforcement – but a well-designed one will be the opposite, focusing on key risks, corresponding to the logical flow of an inspection visit, and clarifying requirements for duty holders.

"Smart Inspections" – using all compliance drivers and differentiated tools

As we have seen with the example of the HSE's *Enforcement Policy Statement,* a balanced inspections and enforcement approach involves the targeted use of a range of instruments – "information, and advice, both face to face and in writing", warnings, and an escalating range of sanctions. A really "smart" approach to inspections includes of course this differentiation in dealing with problems found during inspections, and it also consider inspection visits themselves as but one of a range of possible interventions. Not only are inspections primarily targeted at high risk (and, to a lesser extent, medium risk) objects, but there is also an effort to understand which tools and approaches will be effective to achieve improvements in compliance (and, more broadly, in safety) in particular groups of establishments. In its own "risk-intervention" pyramid, the UK BRDO sees the default type of intervention in low-risk establishments as information and guidance[478]. Even in cases where risk is not trivial, but inspections would be ineffective, looking for alternative interventions is essential. Faced with a problem of unsafe practices in mobile food traders (selling on the highway's side) in South West England in the late 2000s, the Local Better Regulation Office (LBRO – BRDO's predecessor) supported local authorities in developing a "Trader Information Pack". The recognition was that inspections would be ineffective anyway, since mobile traders were, by definition, mobile, and there could be no meaningful follow up, long-term interaction etc. Rather, the key issue was seen to be lack of knowledge, and this was actively tackled. This was linked to a voluntary light-touch "certification" scheme, which allowed to

[477] As Badarch and Kagan (1982) already showed, the emphasis on paperwork is not only ineffective in terms of reducing harms, but also tends to provoke resistance in the regulated entities – and it is most frequently practiced by inspectors with limited competence, and agencies with "no discretion" policies.

[478] Internal documentation, unpublished presentations, interviews with management.

identify "better practices" mobile traders (and notify consumers about them). This was voluntary, but traders who did not join got more checks, hence there was a clear incentive to take part[479].

Another example of a "smart" approach is the development and roll out of the "Safer Food, Better Business" (SFBB) toolkit[480], which we will discuss in greater detail in the third chapter. The development of the toolkit was a response to the entry into force of the new EU "Hygiene Package", and the approach taken stemmed from the finding that many catering businesses had fundamental problems with compliance because of ignorance or misunderstanding of safety requirements, and that this required an approach based on guidance and compliance promotion. In addition, UK food safety authorities had identified the importance of outreach to the many non-English-speaking professionals working in the country's food industry. One of the experiments leading to this acknowledgement was made in Chinatown by the Westminster City authorities[481]. After finding that non-compliances in restaurants were not only frequent, but not improving after repeated inspections, the Westminster regulatory team attempted to understand why. They found out that chefs mostly did not really understand English well, were not aware of local safety regulations, changed repeatedly, and that an inspection with negative findings resulted in a loss of face that made compliance, if anything, even less likely. The response was to emphasize prior training, and to use the chefs' language as much as possible. Along these lines, the SFBB toolkit exists in 16 languages, those most widespread among chefs working in the UK.

In other words, inspections are not a one-size-fits-all. In some cases, they can be a waste of resources, *even* if risks are not negligible. They need to be the appropriate tool to the problem at hand. If the problem primarily stems from lack of knowledge, then punishment will not help, but even an inspection that is not sanctions-oriented but rather primarily consists of advice and guidance may not be the most efficient or effective. Not the most efficient, because it makes more sense to give the knowledge first, through a lower-cost alternative, rather than sending out an inspector immediately. Not the most effective, because in many cases people will listen better to whom they hold to be their "peers" – and they may not accept inspectors as such (depending on whether there is a history of interaction, what are the prevailing regulatory culture and perceptions etc.). Channelling information and guidance through business associations may be in fact more effective. It is partly in recognition of this fact that the UK BRDO has now expanded the "Primary Authority" scheme to small businesses, through their associations. Under Primary Authority, a business that operates in multiple localities in Britain could request to be assigned a "primary" one, which would audit its operations, make recommendations, and issue guidelines on how to inspect and enforce in a given regulatory area, which would be binding for other local authorities also supervising other premises of this business (costs for this in-depth work are to be borne by the business). The scheme has now been extended so that even small businesses operating in one locality only can benefit from it, through their business association. It is the association that will request a primary local authority, and the authority will then issue guidance on how to operate, and how to inspect and enforce, for this given class of small businesses. The expectation is not only that it will make inspections more transparent and consistent (and more risk-based, as BRDO ensures that only the most competent local authorities can be selected as primary) – but also that this will help spread best practices among small businesses, through the guidance given by their associations[482]. With a similar aim, but different means, Lithuania put in place a system of phone and online consultations, whereby businesses can ask their questions about regulations and how to apply them, and get authoritative answers, which they know they can act upon with no fear of inspectors coming up later with a different interpretation (OECD 2015 b). In short, a

[479] Unpublished presentation by Graham Russell, LBRO CEO (and now BRDO Director).
[480] See the Food Standards Agency portal: http://www.food.gov.uk/business-industry/caterers/sfbb
[481] Short case study by the Chartered Institute of Environmental Health: http://www.cieh.org/library/Knowledge/Food_safety_and_hygiene/Case_studies/Westminster%20CHIP.pdf
[482] See Policy Paper *Primary Authority extension and simplification* (BRDO 2015), available at: https://www.gov.uk/government/publications/primary-authority-extension-and-simplification.

"smart inspection" approach is one that recognizes both the importance of inspections (and the need to conduct them in the most professional, efficient and effective way) but also their limitations – and accordingly uses other tools as well to promote compliance and public welfare.

Having sketched out a picture of "risk-based" and "smart" inspections, which includes targeting resources and interventions based on data and risk analysis, increasing inspectors' professionalism and focus during inspection visits, making enforcement responses proportionate, and using a variety of tools apart from inspections to address the diversity of situations and problems, we will now turn to consider some examples from the practice, and try to understand to what extent applying such approaches is relevant to different countries' situations, whether it is realistic, and what results it appears to produce.

The third part will consider data in greater depth. First, its theoretical and actual limitations in terms of allowing us to capture the effects of inspections and of changes in methods. Then, specifically considering the evidence for the contention that risk-based inspections are more effective and more efficient, i.e. produce better (or constant) public welfare outcomes at constant (or reduced) costs. Finally, we will briefly look at what further work could be undertaken in order to produce better, more conclusive data and findings.