The politics of “grayboxing”: The use of QSAR models for the flexible regulation of chemicals

Brice Laurent, CSI – Mines ParisTech

François Thoreau, CSI – Mines ParisTech[[1]](#footnote-1)\*

**Total word count:** papers can be up to 8000 to 10000 words (for now 9000)

**Abstract**

150 words max

**Keywords: QSAR, regulation,**

**Funding**

This work was supported by CSI, ANR… any mention.

**Introduction**

Regulatory bodies make increasing use of computing techniques, models and simulations, for instance so as to predict the toxicity of chemical substances. Such predictions can enable a regulatory decision, typically a market authorization. Therefore, it is important to grasp both the empirical practices of producing and using models, and their political meaning. At stake here is the possibility of identifying the characteristics of a way of governing anticipations that would be based on models.

In this paper, we engage in such an exploration by focusing on a range of models developed to predict the potential risks of chemicals. These models, called “Quantitative Structure-Activities Relationship” (QSAR), are based on statistical correlations between a set of physicochemical descriptors which characterize a substance (e.g. chemical composition, crystalline structures…) and its biological activity, which includes its potential toxicity. In other words, QSAR models rest with the hypothesis that relevant knowledge regarding the toxicity of a chemical can be inferred from its very structure. QSAR models are developed using a limited number of substances that serve as reference points, so that the properties of other chemicals could later be predicted by the model, according to their proximities to the reference points. QSAR models have been promoted by regulatory agencies for over twenty years, but have been recently gaining momentum in Europe, in the wake of the REACH regulation. As the regulation on chemicals is becoming more constraining for private companies, usual experimental approaches raise many concerns (which are lengthy, costly and often requires animal testing).

Based on the analysis of the relevant documentation, interviews with scientists and regulators, as well as the ethnographic observation of a recent research project attempting to develop QSAR models for nanomaterials, this paper provides the preliminary elements for describing the mode of governing anticipation that emerges from the use of such models. We contend that QSAR approaches offer empirical examples to identify a mode of governing anticipation based on flexibility, understood at the epistemic and political levels — i.e. the ability of the regulatory frame to adapt and adjust to tentative categories, but also to test new ways of describing and defining chemicals based on modeling techniques.

Through the discussion of these quantitative risk evaluation and management tools, we propose to examine the politics of modeling as it plays out in European institutions. We analyze it using an approach based on Science and Technology Studies (STS), and particularly a range of scholarly works that have shown the analytical interest of dynamics of coproduction (Jasanoff, 2004). In our case, what is coproduced are the regulatory categories defining the entities expected to be governed (i.e. groups of chemicals), the ways of producing knowledge about them, and the political order whereby powers and modes of action are distributed across public bodies and private companies.

It has been long shown in the field of STS that the very definition of technical entities in regulatory bodies has both epistemological and political consequences. For instance, Sheila Jasanoff has clearly demonstrated that the *description* of GMOs goes along with *prescriptions* about how they should be regulated (Jasanoff 2004; see also Latour 2015). That way, a claim about what the physical world *is* also entails to some extent the appropriate policy choices to undertake. Such a “constitutional work” is particularly visible when considering the operations through “objects of government” (such as regulatory categories describing GMOs or chemicals) are constructed (Lezaun, 2006), and is currently carried out about chemicals within the framework of the REACH regulation at the European level.

Analyzing “constitutional work” requires examining the making of entities expected to be regulated (such as categories of chemicals) or ways of regulating them (such as methods for evaluation or control) and the politics at stake in intervening in these processes. Ultimately, we show that the use of QSAR rewrites basic principles at the heart of the traditional government of risks, such as the definition of stable regulatory categories, the identification of stable validation procedures, and the construction of technical black boxes. Instead, we show that QSAR goes with a government of anticipation based on flexibility (of categories, of tools and methods, of administrative formats). Thus, the discussion of the use of models for regulation making that we propose in this paper offers insight into the analysis of flexibility as an engine for both the production of convincing scientific evidence and legitimate regulatory decisions. Eventually, it sheds light on a constitutional order based on the intermingled roles of public bodies, international organizations, and private actors, the importance of procedural tools, and the permanent redefinitions of objects of government.

The method we follow is to examine the making of QSAR models and the discussions about them in international and European institutions. We use observations from a research project we were involved in, which sought to develop QSAR models for nanomaterials. Our objective is not primarily ethnographic though. We want to build on these observations to reflect on the particular negotiation of public objectivity at play with QSAR modeling: what kind of public proofs does it produce? For what constitutional arrangements? We begin by demonstrating that QSAR modeling implies engaging in a subtle play with accuracy (section 1). This has consequences about the type of objects of government that it makes it possible to define and to regulate. We will show in the second part of this paper that these objects are flexible, as are the methods of validation. Eventually, we discuss the politics of black-boxing and the politics of transparency that flexibility entails (section 3).

**1. Constructing models accurate enough**

***1.1. Choosing descriptors and endpoints***

How to develop a QSAR model? In the project we followed as in other QSAR development projects, the first step was to choose a series of reference substances. These substances were then characterized using “descriptors” – such as size, specific surface area, solubility, etc. and experimental tests are conducted to measure specific “endpoints” – such as cell toxicity, skin irritation, etc. It is understood that neither all the possible descriptors nor all the possible endpoints can be encompassed in a single analysis. Once the two data sets (that of descriptors and that of endpoints) are available, the model itself can be constructed. Its primary objective consists in identifying statistical correlations between the two data sets; where does the description of a substance match with some of its biological effects? Once the model based on these correlations has been crafted, one can thus predict the endpoints for substances that are different from the reference substances, but share some descriptors in common.

Each step in the development of QSAR model (choosing reference materials, choosing descriptors and endpoints, identifying correlation) requires important choices. When examining how these choices are made, the case of nanomaterials can serve as a magnifying empirical lens since all the issues related to the construction of categories (and the high hopes that they carry) are present, only in an even more pressing manner. There are indeed many uncertainties about how best to describe nanomaterials in way that could come to grasp the connections between potential descriptors and endpoints. An official at ECHA describes the nanomaterials situation as follows:

*Because on the top of everything which I told you already, how complex the hazard assessment is by nature, you have to add one additional element that is how size matters, which means how the size of the particles are affecting the toxic properties. (…) The question itself is very complex and then you have to take into account many additional factors on top of everything else. And some of those factors we don’t even have yet fully studied. For example do you know what – how exactly – what is the ratio or the – how the size of a certain particle will for example affect the membrane penetration interaction and so on? We don’t have a direct study which can – we know more or less some thresholds, yeah that something which is smaller than a certain size will easily penetrate the membranes than if something is bigger maybe it will not penetrate the membranes but it can be still by active mechanism uptake and so on, but this is only – we don’t have yet full information to really properly model it.* (interview, ECHA)

This quote is telling, as it clearly states the unease shared by regulators and scientists about how to choose appropriate descriptors for nanomaterials – “appropriate” to the extent to which they would have important consequences for their hazard profile.

During the project we followed, the selection of descriptors was a crucial and much discussed step. The following discussion (between A, B and C, three members of the project) is about whether or not to quantify the form of the substances being used to build the model, and then about what criteria to select in order to differentiate among substances:

1. *Descriptors are not all quantitative… how will we do for the shape of substances?*
2. *So far, what I’ve done is that I have typed the number for each dimension. So if I see “first dimension equals 6”; “second dimension equals 6”; “third dimension equals 300”, I know that it’s a little stick, shaped as a cylinder. (…) Because all our particles have cylindrical symmetry.*
3. *But you could also do, “if it’s a sphere then 1”, “if it’s a cylinder 2”, “3 is a lump”, etc.*
4. *Right, I could separate among all those… Well, what we need to differentiate is among those that are agglomerated or not. (…) There are three or four shapes that we feel like separating, when looking at the pictures.*
5. *We could differentiate among 4 types: isotropic isolated nanoparticles, isolated sticks, isolated bars, and formed aggregates. (…)*
6. *Then there is an ambiguity with boehmite, because boehmite is really bars. But we see sticks, because the bars are superposing themselves – like tiles. Somehow it’s bars and sticks in the same time*
7. *Yeah right, you could do both… but then the question is “what does the cell see?”. And for me, the cell sees sticks. (…) We just take the situation according to the cellular cell, and then it’s not bars. I agree that for a chemist, it’s bars.*
8. *What the chemist sees, and what the biologist sees…*
9. *But there’s no truth in itself here, we choose descriptors from the viewpoint of the cell…*
10. *That’s why when you look at the OECD descriptors, some of them are from the viewpoint of the environment, or from the viewpoint of the river.*

This somewhat long dialogue offers a window into the practical process through which developers of QSAR models choose descriptors. Here, the descriptors being discussed are related to the “shape” of the substances, and what various shapes scientists “feel like separating” among one another, so that a substance on which the model will be used will be described as “particles”, “bars”, “aggregates”… Then the question relates to the number and type of these descriptors of shape. Looking into how this question is dealt with, one can make two remarks.

First, as seen from this exchange, isolating descriptors is a process based on a variety of inputs, including references to guidelines produced by international organizations, considerations about what will make a difference in toxicology effects, and expectations about the potential effects on potential endpoints.

Second, the choice of descriptors is tightly connected to the choice of endpoints. The later part of the dialogue above is about the “viewpoint” of the cell, the environment or the river. If the endpoint is cell toxicity (as it is in the previous excerpt), then the descriptor has to be chosen “from the viewpoint of the cell”. If the endpoint is aquatic toxicity, then the viewpoint will be that of the river. Accordingly, the choice of appropriate descriptors is tightly connected to the potential endpoints one need the model to provide, themselves directly related to regulatory constraints (are the required tests related to cell toxicity? Or to environmental toxicity in aquatic environment?). These two phenomena have consequences. First, the list of descriptors might vary greatly among QSAR models. Second, there is a fundamental uncertainty about the appropriate choice of them, and, consequently, about the categories emerging from the grouping of substances according to descriptors. Getting back to the dialogue above, the project might lead to group substances according to their shapes as “bars” or “sticks”, yet will do so in the context of an interrogation about cell toxicity.

***1.2. Negotiating accuracy***

The selection of reference materials, descriptors and endpoints need to be completed by the exploration of statistical correlations between descriptors and endpoints. This is an iterative process, and yet another step during which the list of descriptors might vary. A person in charge of using QSAR for a French pharmaceutical company explained during an interview that the process of building statistical correlation (that is, the model itself) was characterized by “trials and errors” (she used this expression). If she observed “no answer” from a series of descriptors, that is, that they did not impact the value of the endpoints in statistically significant ways, then she would deduce that they were not relevant, would eliminate them, thereby reducing an initial long list to just a few parameters.

That the choice of descriptors is refined as the model is developed also means that defining the domain of applicability of the model itself is also based on trials and errors. For the objective of a QSAR development project such as the one we studied is to craft the model using known substances in order to enable prediction of potential toxicological properties for other substances, according to their proximity to the ones used to construct the model. For instance, the project we studied used among other substances various kinds of boehmites, working on the hypothesis that these different varieties would offer sufficient information (for instance related to the effects of shape) to situate other nanomaterials according to their proximity to the reference substances.

Yet two problems might arise at this point. The first one is called *overfitting* by QSAR specialists. It means that the model is so tailored to the substances being used to construct it that it is unable to provide any significant information about any other substance. In a case of overfitting, any substance that is different from those used to produce the statistical correlation would be too different for the model to perform. In order to avoid overfitting, QSAR specialists need to build statistical correlations that are *not too accurate*, in order for the model to be usable for new entry data. Overfitting requires that one use a limited number of descriptors so that other chemicals can fit within the model. Yet this raises a second problem, namely that of using too few descriptors for the model to build significant statistical correlation, i. e. *underfitting*. For a correlation to arise, one needs a minimal (yet impossible to know in advance) number of descriptors, various enough for statistical relationships to emerge.

Avoiding the problem of overfitting and that of non relevance requires that QSAR practitioners, such as those we interviewed, proceed with caution and re-invent their methods for each dataset of chemicals used to build models. For them, the objective is to build models *accurate enough*, that is able to formulate reliable predictions for new chemicals which implies they must be not too accurate. To do so, they might process by trials and errors about the list of descriptors, and also about the software tools they use to build correlation. QSAR practitioners we interviewed indeed stated that they might use a variety of software, and adapt their uses according to the responses they get and the correlation they manage to produce.

Thus, in QSAR modeling, accuracy is negotiated. As sociologists of science and technology have demonstrated, constructing accuracy is part and parcel of the making of technological systems, and impacts / is impacted by the larger choices about their objectives and modes of functioning (MacKenzie, 1993). In this particular case, accuracy is permanently negotiated, yet in a way that never aims to construct it as a settled entity. Models need to be “accurate, yet not too accurate”. What does that mean for the government of risks? What does it mean in practice when regulatory decisions have to be made? A way of answering these questions is to analyze the construction of the entities which public regulation hopes to grasp. Governing technical objects require that public bodies craft “objects of government”. In Javier Lezaun’s expression, objects of government are entities constructed through regulatory interventions in order to make it possible to identify and control technical objects (Lezaun, 2006). The next section examines how QSAR models result in the construction of flexible categories of substances, as well as flexible methods of validation.

**3. Flexible objects of government**

***3.1. Flexible categories***

QSAR models impact the construction of categories. The core of model-making is about selecting reference materials and positioning others according to the value of their descriptors. As a result, crafting and using QSAR models implies to group chemicals. Grouping is itself an important task in the regulation of chemicals.

Within REACH, companies need to submit a dossier for each chemical substance they produce or use, in which they demonstrate their ability to evaluate and control potential hazards. For a company putting a chemical on the European market, the question is whether or not the chemicals it aims to register is equivalent to an existing one. For the European public administration, particularly the Helsinki-based European Chemical Agency (ECHA) in charge of the management of REACH, the problem is to evaluate whether or not companies can identify a given chemical to others about which toxicological data already exist. Here, the problem is to group chemicals in categories characterized by similar levels of risks.

Facing the complexities and conflicts that arise while drawing categories, the European Commission has so far adopted an approach of “regulatory precaution” based on a case-by-case approach. Under such an approach, chemicals are considered in isolation one from the others, therefore specific toxicological data should be generated anew. This leads to the endless examination of cases and subsequent multiplication of them, since an infinite variety of parameters can be used to differentiate among substances, e.g. atomic composition, crystalline structure, optic properties, etc. (Boullier and Laurent 2015).

The way QSAR models could overcome such limitations is made particularly salient in the case of nanomaterials. So far this regulatory case has been marked by a twofold shortcoming, i.e. case-by-case vs. arbitrary categories. On the one hand, actual regulatory authorities would endorse the EC’s limited case-by-case approach we just discussed. But on the other hand, the European Parliament tends to propose broad definitions. For example, the European Parliament added an amendment to the 2011 cosmetic regulation whereby it introduced mandatory labeling of cosmetics containing nanomaterials, defined by an arbitrary size limit (set between 1 and 100 nm). Doing so, it would elude any possibility to come to terms with the peculiarities of the materials themselves.

The project in which we have been involved was explicitly designed as a way of escaping this quandary through the use of QSAR models and to overcome this twofold limitation. The aim of the project was to build a predictive QSAR model for nanomaterials, so that it could be possible to group them in several groups, each comprising substances with similar risk profiles. By defining more precisely “profiles” of risk, it would become possible to generate new categories which cut across the whole (made of general definitions) but which also escape the form of nominalism at play when dealing with one single particle at a time (case-by-case).

QSAR methods propose to do such grouping operation, based on physical or chemical descriptors, and associated expected properties. Andrew Worth, a specialist of QSAR methods at the Joint Research Center (JRC) of the European Commission, explains it as follows:

*According to Chapter 3 of the OECD Manual for Investigation of HPV chemicals, a chemical category is defined as:*

*. . .a group of chemicals whose physicochemical and toxicological properties are likely to be similar or follow a regular pattern as a result of structural similarity. These structural similarities may create a predictable pattern in any or all of the following parameters: physicochemical properties, environmental fate and environmental effects, and human health effects.*

*One particular way of expressing the “structural similarity” between different chemicals is to define a structural fragment (SAR) that the chemicals have in common, and a particular way of expressing a regular or predictable pattern is to define a QSAR that links the different chemicals on the basis of physicochemical properties. It therefore follows that SARs and QSARs have a role to play in the formation and rationalisation of chemical categories.*  (Worth et al., 2004: 340)

Accordingly, QSAR methods would allow to group chemicals according to common characteristics that would generate similar physicochemical properties – including those linked to potential hazards. As such, these methods offer ways of grouping chemicals without endlessly separating among them (on the one hand), or creating general and arbitrary criteria (on the other hand).

In the same paper, Andrew Worth states that “however, there is no consensus on how exactly (Q)SARs should be used to define and justify chemical categories.” Andrew Worth’s remark that there is no consensus on the use of QSAR for defining chemical categories is not surprising considering the situatedness of the choices to make (what descriptors? What endpoints? What models?), all depending on regulatory options, available scientific data, and on the personal experience of the people involved in the production of models. This flexibility in the production of QSAR models directly impacts the potential future categories of chemicals based on QSARs. The variety and situatedness of the choices to make imply that categories might be permanently reworked and are, at any case, extremely flexible.

One can wonder whether this flexibility is a first step in the development of models that would ensure the construct of robust categories basing stable objects of government. In other words, are we observing the early stages of a process that will eventually be similar to other regulatory approaches to toxicological risks?

Part of the answer to these questions can be found in the descriptions of the previous section. The choices of descriptors have to be reworked for each model, and depend on endpoints that hinge upon regulatory choices. The construction of significant statistical correlation that could be used to predict potential hazards of chemicals not used to build the models themselves require a subtle play with accuracy that has to be re-experimented each time. Accordingly, the flexibility of category making in an inherent characteristic of QSAR approaches. As such, there is no hope of stabilizing categories once and for all using these models.

The flexibility of categories has important consequences for the government of risks. Contrary to a situation in which objects of government can be settled for regulatory action to take place (e.g. deciding about labeling, control or authorization), the making of categories is in itself a process open for continuous regulatory intervention. This echoes a reflection undertaken by Alberto Cambrosio and Peter Keating about the regulation of biomedical research. Contrasting a situation of “mechanical objectivity”, in which technical instruments can act as providers of accepted upon descriptions of the world, Cambrosio and Keating speak of “regulatory objectivity”, in which public and private institutions need to agree on procedures according to which various regulatory entities can be crafted (Cambrosio et al., 2006). QSAR models do not act as unproblematic black boxes that could provide stable descriptions of the chemical world. For using them as regulatory instruments, agreeing on procedures indeed seems to be an interesting route. Processes of validation of these models might be such procedures.

***3.2. Flexible validation***

Models, in QSAR approaches as in other fields, need to be validated to convince scientific and regulatory audiences of their predictive value. In the case of QSAR, the validation of models has been subjected to intense scrutiny, since it is a condition for models to be used as basis for regulatory decisions (such as restrictions on the circulation of certain categories of chemicals on markets). That the discussion about validation processes has been lasting for years and is still ongoing at the time of writing is significant of an inherent difficult in stabilizing validation criteria. QSAR specialist Andrew Worth explained in a 2004 paper that one could isolate “principles” of validation that could be standardized from “procedures” that would need to be flexible and adapted to each model:

*It is useful to distinguish between the principles of validation, which should ideally be accepted at an international level, and different procedures of validation, which provide different ways of implementing the principles in a practical process. In fact, it is widely agreed that there should be some flexibility in the validation process, so that individual validation exercises can be tailor-made to take into consideration the state of development of individual tests or models.* (Worth et al., 2004: 3)

The distinction Worth proposes here is interesting, since it suggests re-introducing a boundary between standardized “principles” of validation, and adaptable “procedures”, the former being produced through international (possibly “technical”) agreement, and the latter being crafted according to local scientific and regulatory specificities.

The initiatives undertaken at the international level seem to follow Andrew Worth’s 2004 suggestion. At the OECD, the recognition that there were significant variations across countries about the use of QSAR came early, and raised the issue of the harmonization of validation approaches:

*The regulatory use of (…) (Q)SARs varies considerably among OECD member countries, and even between different agencies within the same member country. This is partly due to different regulatory frameworks, which impose different requirements and work under different constraints, but also because an internationally harmonised conceptual framework for assessing (Q)SARs has been lacking. The lack of such a framework led to the widespread recognition of the need for an internationally- agreed set of principles for (Q)SAR validation. The development of a set of agreed principles was considered important, not only to provide regulatory bodies with a scientific basis for making decisions on the acceptability (or otherwise) of data generated by (Q)SARs, but also to promote the mutual acceptance of (Q)SAR models by improving the transparency and consistency of QSAR reporting.*(OECD, 2007: 15)

Harmonizing QSAR validation processes eventually took the form of general principles agreed upon at the OECD. These principles are the following:

“To facilitate the consideration of a (Q)SAR model for regulatory purposes, it should be associated with the following information:

1. a defined endpoint;

2. an unambiguous algorithm;

3. a defined domain of applicability;

4. appropriate measures of goodness-of-fit, robustness and predictivity;

5. a mechanistic interpretation, if possible” (OECD 2007: 14).

The production of such principles within the OECD seems to follow Andrew Worth’s suggestion that “principles of validation” should be agreed upon “at the international level”. These principles indeed offer a way of ensuring international agreement about QSAR validation processes. Yet they had to do so without entering the domain of regulation making, which was that of sovereign policy choices, outside the scope of OECD intervention. For the OECD intervention to be acceptable, as that of a provider of “apolitical” international expertise, QSAR validation principles had to be framed in a general way. Consequently, no considerations about relevant endpoints could be acceptable, since the choice of endpoints is directly related to regulatory decisions. This consideration is made explicit in the OECD’s self-description of its principles of validation, meant to ensure a level of generality that would not cross the boundary between international expertise and sovereign regulatory choices. Consequently, these principles are not enough to practically validate models, and might be ambiguous (this is particularly the case for the 5th principle, “mechanistic interpretation, if possible”, which attempts at re-introducing a chain of causality that is not statistical (Thoreau, forthcoming)).

At the OECD, the five principles are defined in such a way that they do not enter the domain of regulation making, characterized notably by the definition of endpoints. Within the European institutions, the separation plays out somewhat differently, since endpoints are defined within the European regulation. Consequently, validation principles within the European institutions can be more operational than the OECD ones. Yet more difficulties arise. Validating QSAR models can be carried out processing the data that have been used to construct the statistical correlations (this is described as “internal validation”), or other data (e.g. chemicals of known risks, on which the model will be run, and its predictions checked against the known risks of the tested chemicals). The latter approach is called “external validation” and is deemed more robust for regulatory choice by QSAR specialists (Gramatica, 2007). Yet external validation also requires additional data, and additional testing to check whether the predictions according to the model are correct, and yet another validation process for the choice and use of these additional data.

When the diversity and permanent evolution of statistical tools are added to the picture, the conclusion, drawn by Andrew Worth himself, is that the validation process cannot be “set in stone”:

*There should be nothing to fear from this process, since no conclusion on the validity of an experimental test or a (Q)SAR model is ever set permanently in stone—scientific and technical developments should always be taken into account. The question will always be when should the validity of a (Q)SAR (or a test method) be reviewed, either due to an adaptation of the model (test) itself, or because a new assessment (e.g. statistical) method is developed, or because new information (e.g., test data) becomes available.* (Worth et al., 2004: 356)

In practical terms, this means that the standardization of validation processes can only take the form of general principles, leaving the practical conduct of validation to the particularities of the regulatory and technical situations at stake. Depending on the type of chemicals and models, internal or external validation processes will be used, and in ways that will differ from one case to the next. Thus, QSAR practitioners and regulators need to re-question the appropriate validation methods for each new situation.

***3.3. Flexible use***

As category-making based on QSAR is inherently flexible, so is the validation process. Accordingly, the 2008 guidelines proposed by ECHA about the grouping of chemicals based on QSAR do not propose to standardize the validation methodologies, but mean them to be permanently discussed by the actors involved, and regularly adapted. Consequently, the use of QSAR for regulatory decision cannot take the form of the mobilization of a ready-made instrument about which there could be unambiguous consensus on its scientific validity:

*Even though computer-based estimation tools are becoming increasingly available, these tools are intended to facilitate the process of (Q)SAR acceptance and cannot substitute the need for expert judgement and dialogue between industry and authorities. The use of (Q)SAR predictions in an automatic way, without considering validation results, regulatory purpose and use of WoE [Weight of Evidence] judgements is not recommended. Having said that, on the basis of current experience, it is difficult to give detailed guidance on how to use (Q)SAR estimates for regulatory purposes. Indeed, it is debatable to what extent it will be possible to codify accepted practise in terms of rules-of-thumb, although some attempts must be made along these lines.*

*The approach proposed is that experience in the regulatory use of non-testing data should be obtained by following a learning-by-doing approach, with the learnings being documented as examples for reference purposes. In this way the possibilities for enhanced use of non-testing methods in general under REACH will be optimised whilst avoiding long bureaucratic and formal adoption schemes* (ECHA, 2008: 26-27)

A quote such as the previous one offers a striking difference from what is expected from expert and regulatory institutions in charge of the government of risk. The objective here is to maintain enough flexibility so as to adjust any regulatory process to the particular situation at stake, knowing that no standardization of the method can be in sight. Yet, and as STS scholars have shown, producing public proofs for the public regulation has traditionally relied on boundary-making processes, such as the ones through which “risk assessment” and “risk management” are isolated from one another, both in the conduct of scientific work and in the institutional organization of expert agencies and regulatory institutions (Jasanoff, 1990). These boundaries are permanently at stake, not less because of the inherent value choices at the heart of risk assessment – regarding, for instance, the working hypothesis, the choices of priorities of studies, or the management of scientific uncertainties[[2]](#footnote-2). Yet they are the epistemic and organizational basis for grounding both scientific objectivity and democratic legitimacy. They have constitutional effects, at both the level of the constitution of objects of government and that of the allocation of power between public bodies. They rely on a form of mechanical objectivity whereby instruments can stabilize descriptions of the technical world. QSAR models radically reframe the basis of such an arrangement. Examining these processes is a way of discussing the politics of flexibility as it appears in QSAR modeling: once the boundary between risk assessment and risk management is no longer the basis for the government of risks, what entities are being crafted (ontological)? What is the allocation of power, the distribution of resources, the ability to act for public and private actors? What is negotiated? Where and by whom? These questions are examined in the next section.

**4. The politics of flexibility**

***4.1 A politics of black-boxing***

As sociologists of science, black-boxing is central in the construction of science (Latour, 1990), in that it allows researchers not to re-open intermediary results and experiments devices each time they are used. For regulators, black-boxing is equally important. By many respects, the constitutional arrangements based on the separation between “assessment” and “management” is based on such a process. Only when black-boxing has been performed will it become possible to claim that “assessment” is free for social and political considerations. By contrast, situations of controversies about risks mean that black-boxes are open, evaluation processes examined, and well cared after boundaries crossed (see examples in Jasanoff, 1993). In the case of QSAR modeling, the flexibility of categories and evaluation processes make black-boxing an interesting process for regulators. Indeed, a way of reducing variety in the use of models could be provided by black-boxing processes – whereby some elements (such as calculation tools) are agreed upon and can circulate.

Such black boxes can then be resources for both private and public actors. They make it possible to develop new markets (e.g. of software developed by consultancies). They offer ready-made tools for regulators to use. The OECD “toolbox” can be seen as an attempt at black-boxing. The OECD has been active in the production of expertise about chemical testing since the early 1980s, thereby defining itself as an expert international institution able, as a legal scholar puts it, “to partition the technical aspects from the political” (Salzman, 2005: 203). The QSAR toolbox could be situated within this overall objective. It is being developed in coordination with ECHA and was originally conceived as a way of delegating the construction of an instrument that could systematize the grouping of chemicals based on statistical correlations, and according to endpoints that would be chosen by the user of the toolbox. It would then offer yet another model, but possibly more robust and more widely accepted that other ones. The implication of ECHA in the development of the toolbox was described as such by an official during an interview:

*(In 2008…), ECHA decided to support OECD in the development of the project. So after a proof of concept phase, ECHA funded the second phase of development and now we are funding already a third phase of development which is for the next four years. So basically we have been investing for six, almost seven years in this tool already. (…) we are not only investing money but we are investing our expertise and our expectations as well, which means that we are also sharing our experience with QSAR and (… exploring) how we could make the Toolbox nicer, more robust, easier and so on.* (interview, ECHA)

In doing so, the delegation of the production of the toolbox to the OECD is a way for ECHA to ensure that a component of a regulatory approach is developed independently from regulatory choices, so that it could travel easily across various cases. This is made explicit in the internal organization of work at the OECD. The same person from ECHA re-stated in a way the importance of the science/policy boundary, in order for a technical black-box to be developed independently from policy considerations, as he made the following comment:

*So this is something which we are doing, but (…) this is not like we are making decisions together with our colleagues from OECD, no we have – the project itself has a Management Group which is responsible for deciding which kind of models are good enough to be included in the Toolbox.* (interview, ECHA)

The development of the toolbox was conducted step by step. The latest version of the toolbox was released in December 2014, with, according to the person in charge of the toolbox at the OECD, “all kinds of new functionalities”[[3]](#footnote-3). When he mentioned these “new functionalities”, this person was referring at first to design choices making the toolbox more “user-friendly”. This evolution is more than just about the appearance of the tool. It is significant of an iterative process inherent to the development of QSAR models. Making the toolbox even more robust means providing the toolbox with data. The more data the toolbox can process, the more robust the model is. Hence the successive versions of the toolbox rely on experimental data provided by various actors. ECHA provides data provided by the companies that register their substances; other actors such as private companies or organizations developing QSAR provide more precise date, targeting for instance particular endpoints (such as carcinogenic effects). The permanent refinement of the toolbox according to the data it is fed with is a never-ending process – and inherently so since there is an infinite variation of substances, descriptors and potential endpoints.

This requires a permanent revision of the data with which the toolbox is fed. This is undertaken by the members of the toolbox management group, which includes, among others[[4]](#footnote-4), experts from member countries regulatory authorities, and is not a straightforward process, as explained during an interview by the person in charge of the toolbox at the OECD:

*Okay so basically first – always the decision about inclusion to the Toolbox, if some database should be included or not to the Toolbox is done by the Management Group. So basically if you want to include something to the Toolbox, you need to do the proposal to the Management Board. (…) So basically you need to describe all the facts about this data, how the data was generated or collected and so on and based on that, the Project Team is trying to assess more or less the database, so see how it looks (…) and see if this data is consistent and what might be the problems in integration with the Toolbox and makes the recommendation to the Management Board. And then the Management Board takes into account the original proposal plus Project Team recommendations in deciding, “Okay if we want to include it or not”. (…) So I think currently we have like around 40 different databases included in the Toolbox.*

Thus, feeding the toolbox with new data is not only a never-ending process, it is also a long one, implying a permanent assessment of proposed dataset, and a permanent enrolment of new users who could provide more data. This latter point helps connecting the seemingly external issue of user-friendliness with the development of the content of the toolbox itself. As more and more datasets are needed in order to successfully develop the toolbox, as larger datasets are deemed to provide more refined results, then it is important to make the toolbox usable for more users, by for instance, proposing them to select endpoints of potential interest and providing them with relevant descriptors.

One can then grasp an important characteristic of the development of a would-be black-box such as the OECD toolbox. Such a development implies enrolling more and more users, as a way of ensuring both the collective legitimacy and the technical validity of the instrument. This enrolment process is a constitutive part of the toolbox and not a primary phase before stabilization. In other words, the toolbox cannot provide a stable public proof of its scientific validity and regulatory acceptability, but is based on the permanent re-working of the public proof.

A consequence of this (and of the fact that the toolbox is just one possible instruments among many others, such as private software, for the grouping of chemicals and the anticipation of their risks), is that no general black-boxing is possible. Within the European institution, this is formulated as such:

*The process of (Q)SAR acceptance under REACH will involve initial acceptance by industry and subsequent evaluation by the authorities, on a case-by-case basis. It is not foreseen that there will be a formal adoption process, in the same way that test methods are currently adopted in the EU and OECD. In other words, it is not foreseen that there will be an official, legally binding list of (Q)SAR methods.* (ECHA, 2008: 27)

This quote contrasts the test methods developed by the EU and the OECD and regularly used to conduct toxicity assessment. No such consensual production exists for QSAR, and this is not a matter of waiting long enough for standardization to be possible. This is inherent to QSAR itself, and implies that rather than relying on transferable black-boxes (such as the test methodologies produced by the OECD about experimental toxicity), the model-based government of risks is based on the permanent re-working of public proofs. Consequently, black-boxing can never be entirely settled. This means that the production of black boxes can take various forms. It is open for public and private interventions. It results in permanent negotiations about the circulations and uses black boxes.

**4.2. *A politics of transparency***

Even if black-boxes can be and are produced, for instance by private companies providing software for specific group of chemicals, the use of such black-boxes by the European public regulators is ambivalent. Since many choices need to be made to construct and use models, then regulators might be wary of too complex instruments that they would be unable to apprehend (regarding e.g. the hypothesis, the domain of applicability, or the statistical methods being used). Even in the case of the internationally agreed-upon OECD QSAR toolbox, black-boxing goes hand in hand with a constant requirement for transparency. An ECHA official put it in the following terms during an interview:

*I think what OECD understood very quickly, (…) that the most important, most critical element for Regulators is the transparency of the model. If you have a very sophisticated statistical model (…) this is not very convincing for Regulators because they don’t exactly know what was exactly the training set which you used to train those networks and even if you see that they are performing very well on your test validation set, it doesn’t mean that they will perform equally good on the new substance which are out of the validation set. And this is the basic problem of all those advanced QSARs, that they are not so transparent because they are very complex and regulators have always this problem in understanding what will the logic behind the tool? What kind of features were driving predictions?* (interview ECHA)

The successive versions of the toolbox reflect this requirement for transparency. For instance, the newest version of the toolbox makes it possible for users to have information about the datasets:

*What we’re also going to develop in the new version is to have a kind of reliability score related to the database and the profile so that at least they are all well documented so the regulators can decide that we are very transparent on how these databases or profiles are constructed, what kind of chemicals have been used to develop – which are included in the database. So if you go to the Toolbox you also have an “about” section. You select a database and click on the “about” section then you will get information on the database.* (interview, OECD)

The concern for transparency in QSAR modeling is not limited to the OECD toolbox. In fact, the issue is even deeper in for others tools, such as software developed by private companies and used by companies in ways that are not always easily understandable by regulators assessing registration dossiers. The use of QSAR for regulation making is indeed characterized by a multiplicity of would-be black-boxes produced by public and private actors developing software, based on different statistical approaches. Although guidance documents have been produced by public bodies such as ECHA, and although attempts at producing common instruments such as the OECD toolbox have been undertaken, the situation of diversity of instruments still remains. Consequently, public agencies are in a permanent need for re-opening the black-boxes used by companies. Consider for instance how members of the French public agency for environmental safety describe their roles in assessing the use of QSAR models by companies:

*- And I think that the stake for us is to identify the limits and confront the industrialists. (…) If we are not able to deconstruct the reasoning and know what there is in black boxes, then we can’t argue with what companies propose! We can’t say that we don’t accept because we would have checked the domain of application, or whatever. That’s why we need internal competencies for that… -… for a counter-expertise really.* (interview, ANSES)

Thus, opening the black-box of models is a permanent requirement of regulators, which makes them prefer hybrid approaches combining experimental and modeled data rather than unique modeling solution based on sophisticated, black-boxed tool:

*Yes and this is something which is important for regulators because if I will know that for example the prediction is backed up by some solid hypothesis which is confirmed by for example different in vitro observations or other observations in vitro from similar substances, this is for me something much more important than just predictions generated by super duper fancy logic, for example neural networks. And another important point is really transparency. Regulators are not looking for the tool which will give you the smallest possible error in predicting something on your validation set; regulators are more keen on something which they can understand how it works and they can extrapolate it to the normal – their own experience. It’s even easier to accept the tool which gives you some error, like for example a few units plus or minus, but you know that this is really more or less what’s going on and this sounds reasonably good, rather than using some very advanced mathematical model which you cannot really follow and you don’t even know exactly how those features have been generated by the model.* (Interview, ECHA)

Thus, the regulatory use of models plays on a subtle balance between the mobilization of black-boxed tools and a call for transparency that makes regulators refrain from accepting too sophisticated models that could not be easily deconstructed. In practical terms, this also means that regulators need to craft administrative procedures that would allow them to gather information about how QSAR models have been developed and used.

An example of such procedures is the “QSAR model reporting format”, described as follows by Andrew Worth during an interview:

*Documenting models in a consistent way (…) doesn’t go in to the minutiae of model development and data curation and so forth, it – in a sense it describes what you get at the end of the model building process and it doesn’t describe all of the steps during the model building process and if you like the quality assurance that’s gone in to that so if you’re thinking of something like good laboratory practices for in vitro methods, that’s – it’s not just documenting what comes out of the pipeline, it’s documenting the procedure and that you know – the traceability of that procedure and the QSAR validation – it’s called the “QSAR Model Reporting Format” and it complies with the Validation Principles, really just documents what comes out at the end and it doesn’t go in to all the details of how you’ve done the model and how you check the quality of data and so forth.* (Interview, A. Worth)

This quote clearly shows that the requirement for transparency about models does not mean that all possible information about QSAR models will be provided to regulators. Rather, what matters to them is to identify the few pieces of information they need to assess the type of validation. This concern about the appropriate reporting format is telling is not anecdotal. It is part of the consequences of using QSAR modeling for the relationships between public bodies and private actors. Faced with the proliferation of would-be black boxes they cannot completely take for granted, public experts face an asymmetry of resources, requiring that they craft procedures whereby they can gather enough information to make regulatory choices. For private companies, the submission of dossiers becomes more strategic than ever, since the plurality of available models mean that some of them might suit their needs and interests better than others. Consequently, as the production of black boxes results in growing roles for private actors and increasing demands on public bodies in charge of questioning the models used by industries, so the requirement for transparency results in attempts at defining ways of opening up the models, or at least crafting procedures to gather information about them. While QSAR models destabilize the traditional constitutional arrangement at the heart of the public regulation of risk (where risk assessment and risk management are neatly separated), these tools require that the relationships between private companies and public regulators are transformed. New private actors enter the picture (most notably the companies producing models), producers of chemicals need to engage in new strategic activities (choosing relevant models), and public regulators need to re-invent their roles so that they are able to monitor both the construction and the use of models.

**3. Conclusion**

In this paper, we have discussed the ways in which QSAR models propose a mode of governing anticipation based on flexibility. Using these models for the regulation of chemicals indeed implies:

* The construction of flexible categories, as the grouping of chemicals is not settled once and for all, but redrawn according to scientific and regulatory considerations embedded in the very construction of QSAR models;
* The definition of flexible validation processes, as the validation of QSAR models cannot be standardized but re-adapted for each case;
* A process of flexible black-boxing, in which the production of instruments expected to provide public proofs is permanently re-worked, and black-boxes are constantly re-opened by regulators in search for transparency.

The government of anticipation that emerges from these considerations is quite different from the public administrative of chemicals risks based on experimental toxicology. In the latter case, categories are supposed to be defined, and phases of risk assessment and risk management neatly distinguished. In the case of QSAR models, the forms of scientific validation and political agreement are quite different. Flexibility, for that matter, appears as both an epistemic and a political characteristic. We can hypothesize that it describes an emerging mode of government that could be identified on other technical domains as well.

**Acknowledgments**

The authors are grateful to XXX.

**References**

Boullier H. and Laurent B. (2015), “La précaution réglementaire. Un mode européen de gouvernement des objets techniques”, *Politique européenne*, vol. 49, pp. 32-55.

Cambrosio, Alberto, et al. "Regulatory objectivity and the generation and management of evidence in medicine." *Social science & medicine* 63.1 (2006): 189-199.

Gramatica, Paola. "Principles of QSAR models validation: internal and external." *QSAR and Combinatorial Science* 26.5 (2007): 694.

ECHA, 2008, *Guidance on information requirements and chemical safety assessment. Chapter R.6 : QSARs and grouping of chemicals*, European Chemical Agency.

Jasanoff S. (2004), *Designs on Nature: Science and Democracy in Europe and the United States*, Princeton: Princeton University Press.

Jasanoff, Sheila. "Contested boundaries in policy-relevant science." *Social studies of science* 17.2 (1987): 195-230.

Jasanoff, Sheila. *The fifth branch: Science advisers as policymakers*. Harvard University Press, 2009.

Latour B. (2015), *Face à Gaïa. Huit conférences sur le nouveau régime climatique*, Paris: La Découverte / Les Empêcheurs de penser en rond (English translation forthcoming, in the meantime see the series of Gifford Lectures, *Facing Gaïa*. *Six lectures on the political theology of nature*, [www.brunolatour.fr](http://www.brunolatour.fr)).

Lezaun, Javier. "Creating a New Object of Government Making Genetically Modified Organisms Traceable." *Social Studies of Science* 36.4 (2006): 499-531.

OECD (2007) *Guidance document on the validation of (quantitative) structure-activity relationship [(Q)SAR] Models*, ENV/JM/MONO(2007)2.

Salzman J (2005) “Decentralized administrative law in the Organization for Economic Cooperation and Development”. *Law and Contemporary Problems* 68:3-4.

Thoreau, François, forthcoming, “’A mechanistic interpretation, if possible’: how does predictive modeling causality affect environmental regulation?”, accepted for publication in *Big Data & Society*.

Worth, Andrew, T. Hartung and CJ Van Leeuwen, 2004, “The role of ECVAM in the validation of QSARs”, SAR and QSAR in Environmental Research, Vol. 15 (5–6), October–December 2004, pp. 345–358

1. \* Corresponding author: Brice Laurent, Bd Saint-Michel etc. [↑](#footnote-ref-1)
2. Sheila Jasanoff has shown that the 1983 report of the U.S. National Research Council, *Risk Assessment in the federal gouvernment : managing the process*, known as the *Red Book*, established the priority of the risk assessment/risk management boundary, while simulataneously recognizing in explicit ways that risk assessment was a value-ladden process. [↑](#footnote-ref-2)
3. Interview, OECD toolbox team [↑](#footnote-ref-3)
4. While the composition of the group remains secret, it is composed by representatives from national regulatory bodies, industries — esp.from the pharmaceutical domain — and from international institutions. [↑](#footnote-ref-4)