Situated Expert Judgment: QSAR Models and Transparency in the European Regulation of Chemicals

Brice Laurent

MINES ParisTech, PSL Research University, CSI – Centre de Sociologie de l'Innovation, i3 UMR CNRS, France/ brice.laurent@mines-paristech.fr

François Thoreau

Spiral research centre, University of Liege, Belgium

Abstract

This paper discusses the kind of expert judgement demanded by the development of a particular class of models. It analyses the case of 'Quantitative Structure-Activities Relationship' (QSAR) models, used to predict the toxicity of chemical substances, for regulatory and other purposes. We analyse the production of these models, and attempts at standardizing them. We show that neither a technical nor a procedural standardization is possible. As a consequence, QSAR models cannot ground a production of knowledge along the lines of 'mechanical objectivity' or 'regulatory objectivity'. Instead, QSAR models imply that expert judgement is situated, re-worked for each new case, and implies an active intervention of the individual expert. This has important consequences for risk governance based on models. It makes transparency a central concern. It also means that new asymmetries emerge, between companies developing sophisticated models and individual experts in regulatory agencies in charge of assessing these models.

Keywords: expertise, models, objectivity, regulation, transparency

Introduction

Computer simulation and computer modelling are being used to govern a growing share of social activities. A recent evolution has made computer models a tool for evaluating and controlling the health and environmental risks raised by chemicals. Using statistical correlation, models would predict which chemicals are problematic, and complement other risk assessment methods such as *in vitro* or *in vivo* tests. In situations where scientific uncertainty is present, models would provide additional scientific elements to ensure that regulatory decisions are appropriate.

Described as such, it would be tempting to see models as ready-made scientific tools expected to provide objective descriptions of technical entities, for later use in regulatory settings. But what 'objective' means in this context is not self-evident. Works in Science and Technology Studies

(STS) have shown that objectivity is manufactured in various ways, which differs across historical and regulatory contexts, and which directly impacts how expert judgment is conducted (Cambrosio et al., 2006; Cambrosio and Keating, 2009; Daston and Galison, 2007; Jasanoff, 2011). One of the important insights of STS works on objectivity is that the production of objective knowledge implies that the human subjects expected to produce or witness objective knowledge are shaped in particular ways. In regulatory settings, this means that the production of objective knowledge also defines the type of expert judgment at stake.

We follow this inspiration in this paper. We examine the use of models for regulatory purposes by analysing the expert judgment that it entails. We focus on models known as 'Quantitative Structure-Activities Relationship' (QSAR), designed to predict the toxicity of chemical substances. These models are based on statistical correlations between a set of physicochemical descriptors that characterize a substance (e.g. chemical composition, morphology, …) and its biological activity, including its potential toxicity. In other words, QSAR models are based on the hypothesis that relevant knowledge regarding the toxicity of a chemical can be inferred from its structure. Diverse actors are developing QSAR models and produce a multiplicity of different QSAR models for different purposes (Lo Piparo and Worth, 2010). They thus embody the diversity and complexity of foreknowledge used in policy. Like other models in various technical areas, QSAR models are used by policy-makers to inform regulatory decisions. And like other models, they raise a series of uncertainties that have political consequences (see e.g. Edwards, 1999, 2010 about climate modelling).

Our objective in this paper is to analyse the political issues raised by QSAR models, particularly focusing on the ways by which they challenge the practice of public expertise. We argue that QSAR models are empirical entry points to reflect on risk governance based on models, and in particular the type of expert judgment that this approach entails. We demonstrate that the expert judgment that these models require cannot be tied to the use of ready-made technical tools providing stable scientific evaluations (as in situations of 'mechanical objectivity', see Daston and Galison, 2007), nor to procedures and standards framing the appropriate mode of action (as in situations of 'regulatory objectivity', see Cambrosio and Keating, 2009). Instead, QSAR models imply that expert judgement is situated, re-worked for each new case, and implies an active intervention of the individual expert. This has important consequences for risk governance based on models. It makes transparency a central concern. It also means that new asymmetries emerge, between companies developing sophisticated models and individual experts in regulatory agencies in charge of assessing these models.

Echoing Boullier, Demortain and Zeeman who look at the beginnings of QSAR modelling in chemicals regulation at the US Environmental Protection Agency (Boullier et al., 2019), our focus is on European institutions, and how they use or plan to use QSAR models for the regulation of chemicals, possibly by using international standards. Chemicals are regulated in Europe within the REACH regulation (Registration, Evaluation, Authorization and Restriction of Chemicals; see European Commission, 2006). Within this framework, companies have to demonstrate to public expert bodies that they are able to evaluate and manage the risks of the substances they produce. This requirement results in a large number of toxicological tests. The use of computational models could appear as a means to mitigate this trend.

Following an approach undertaken by scholars who have examined the use of models in policy arenas (Edwards, 1999; Heaphy, 2015; Fisher et al., 2010), we examine the making of QSAR models and the debates about them in European institutions, as well as in the Organisation for Economic Cooperation and Development (OECD) where such discussions are held and European actors are involved. We base our reflection on three sets of empirical material. First, we use observations from a research project that developed QSAR models for nanomaterials. This research project involved material scientists and toxicologists, and was conceived as a demonstration of the interest of the QSAR approach for regulatory purposes. We were involved in the project for a year in 2014-2015[1]. We observed research meetings and conducted

interviews with the leaders of the toxicology and materials science teams, as well as with the post-doc researcher and the engineer involved. Second, we examine standardization attempts at the OECD, which were expected to help public bodies evaluate the use of QSAR models by private companies. We use the OECD literature on the topic, as well as two interviews with participants in the OECD working groups. Third, we build on five interviews with experts working in public organisations in charge of evaluating what private companies submit to register the chemicals they produce within the REACH framework. We use this qualitative empirical material to infer how QSAR models are expected to function with risk governance frameworks. Taking inspiration from Science and Technology Studies (Jasanoff, 2004), we discuss the type of technical knowledge that QSAR models are expected to provide, and the expert judgment that this knowledge requires.

The progression of our argument mirrors the list of our empirical sites. First, we situate our approach in a more general debate about expert judgement. We then examine the practices of QSAR model-making and elaborate further about QSAR models as outcomes of trial-and-error processes, unfit for reaching definitive closure, as indicate the efforts coordinated by the OECD in order to standardize processes of validation. We then analyse the "QSAR toolbox" developed by the OECD. The toolbox provides an evolving and flexible tool amenable to public experts uses and appropriations. Lastly, we finish by analysing the consequences of the previous considerations for the works of experts working in public agencies in charge of evaluating the use of QSAR models. We show that QSAR models require expert judgment to be defined in situated ways, and that this situatedness makes transparency a key component of the risk governance framework, in turn producing new asymmetries between public bodies and private companies.

Expert judgment and the problem of QSAR models

Risk governance relies on the ability of public institutions to mobilize technical expertise for decision-making. Expertise has been famously problematized by Sheila Jasanoff (2005) as a 'three-body problem', in that its legitimacy is the outcome of a subtle articulation between a public body organized to deliver expertise, a body of knowledge stable enough to provide grounded facts, and the body of the expert as an individual expected to provide consequential advice (Jasanoff, 2005). This perspective shows that the form of this articulation may vary. Throughout Jasanoff's works, the American case appears as a particularly interesting illustration of the importance of the 'view from nowhere' in defining expert legitimacy. The 'view from nowhere' points to the set of mechanisms whereby expert advice is disconnected from the particularities of its conditions of production, whether related to situated technical choices or to the individualities of the experts themselves.

Problematizing expertise as the outcome of a view from nowhere has consequences for both the organization of public institutions and the type of expert judgment. First, it implies that risk assessment (as an outcome of expert judgment) is carefully separated from risk management (where decisions can be related to particular decision-makers and political stakes). Second, experts ground the legitimacy of their interventions on their ability to ensure a form of 'mechanical objectivity' (Daston and Galison, 2007) whereby instruments can stabilize descriptions of the technical world purified from human intervention. Despite this importance in the organization of American expert bodies, this configuration is only painfully and temporarily stabilized, as regulators themselves acknowledge the inter-relatedness of risk assessment and risk management, experts' political motivations are questioned, and the very ability to operate the view from nowhere in practice is questioned (Hilgartner, 2002; Jasanoff, 1990). When expert judgment is framed as the outcome of the view from nowhere, experts are expected to disappear behind the instruments they mobilize. Mechanical objectivity relies on instruments that can travel in a stable way, and ensure robust fact-making because of their stability. As such, they are black-boxes in the Latourian sense (Latour, 1987). This does not mean that experts as human beings are no longer individuals on their own, but that public institutions

define the legitimacy of their interventions in their ability to make their selves independent from the production of facts.

The debates about European expertise can be read as a variation on the three-body problem of expert legitimacy. They display a pervasive tension between attempts at reproducing the 'view from nowhere' and the political negotiations at the heart of the European regulation of technical objects. Thus, when European institutions responded to food crisis by the creation of a centralized expert body (the European Food Safety Authority) expected to operate independently from political pressure (Demortain, 2009), they were caught in pervasive tensions about whether or not the experts of the agency were actually free from private interests (Vos, 2000), and about whether the agency could provide technical advice expected to ground decisions for all member states (Wickson and Wynne, 2012). The difficulties that an expertise body such as EFSA has encountered can be interpreted as outcomes of a pervasive tension within the European expertise institutions. In the European context, manufacturing expert judgment is directly connected with the negotiations between member states and stakeholders (see: Saurugger, 2002). This makes the call to reproduce the 'view from nowhere' highly problematic, since this configuration might neglect the specificities of the European political landscape and modes of negotiation.

The case of chemicals however seems to provide an illustration of a successful stabilization of European expertise. Within the REACH regulation, the European regulation of chemicals is based on the coordinated action of the European Chemicals Agency (ECHA) and national expert bodies, the former acts as a centralizing body able to leave room for national variations (Boullier, 2016). Techniques through which assessments can be conducted in uncontroversial ways are therefore even more important. Some of them are procedural (as registration dossiers codified in European regulations), while others are based on standardization. Examples of the latter include the methods described in the technical guides published by the ECHA to help operationalize REACH, and standardized testing methodologies produced by the OECD, intended to be technical tools neatly distinguished from the regulatory choices that sovereign members of the international organizations might make (Salzman, 2005: 203).

For all their diversity, these tools have a similar role in the REACH risk governance framework, namely to provide the European experts with stabilized tools able to ensure the technical validity of risk assessment as they examine registration dossiers for chemicals. They serve as instruments through which experts working at ECHA can evaluate the dossiers submitted by private companies as they ask to register the substances they produce. These tools are expected to ensure that the outcome of expert advice only depends on the instruments being used and not on the individuality of the expert conducting the evaluation. Eventually, they allow these experts to separate the technical phase of risk assessment from the political phase of risk management.

QSAR models have been promoted by regulatory agencies for over twenty years, but have recently gained 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. Testing methods are lengthy, costly and often require animal testing. REACH is gradually extended to larger families of materials, which implies that even more tests need to be performed to ensure that chemicals can circulate on the European markets. In this context, QSAR models could appear as an alternative. They could provide knowledge about the potential risks of a given substance without conducting any test. In practice, this means that a company wishing to register a new substance could argue, based on models, that this very substance has a risk profile similar to other substances already registered.

Models could provide an additional resource for European expertise to ensure its technical validity, and its ability to be distinguished from regulatory decisions. However, the ECHA experts do not present QSAR models as technical black boxes that could ground a mechanical objectivity expected to make subjective interventions disappear. Consider the ways in which ECHA

presents QSAR models in one of its 'guidance documents':

> 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)

The contrast with the OECD tests is interesting since the latter are a good illustration of internationally agreed-upon methods that can act as resources for the technical validity of the assessment. By contrast, in regulatory decision making, QSAR models cannot be seen as ready-made instruments with unambiguous consensus on their scientific validity.

The ECHA (2008: 26-27) document quoted above states that the "use of (Q)SAR predictions in an automatic way" is "not recommended". Instead, it asks experts to consider "validation results, regulatory purpose and use of weight of evidence" (ECHA, 2008: 26-27). Rather than offering ready-made instruments providing scientific evidence for the technical phase of risk governance, independently from particular regulatory choices, QSAR models seem to be far from universal acceptance, and to be effectively tied to particular regulatory considerations. The specialists of QSAR modelling whom we met concurred. Many of them saw QSAR models as tools for conducting a preliminary selection of potentially problematic substances (or "screening," as they would say), while being extremely wary of a potential use that would go beyond providing additional evidence to that produced through standardized testing.

Thus, when discussing the use of QSAR models within the ECHA, specialists of the methods explain that:

> Under the coordination of the Chemicals Agency, the regulatory bodies in the EU will then make case-by-case decisions on the acceptability of any (Q)SAR models and estimates used, taking into account the regulatory context and the availability of other information. (Worth et al., 2007: 116)

Such wording seems to imply that QSAR methods are not expected to become black-boxes ready to be used as proof- making devices, but are tied to local conditions of use. This, we contend, prevents expert judgment from relying on a form of mechanical objectivity, and entails new political challenges, in terms of the identity of the actors involved in risk governance, the possibility of publicly controlling them, and the nature of public proof. To understand these challenges, we need to demonstrate that the impossibility to black-box QSAR models is not a mere incidental and preliminary situation before their eventual stabilization, but part of their very nature, and of what makes them of interest to industrial producers and public experts in the first place. To do so, we need to delve into the mechanisms of model-making. This requires that we temporarily leave the world of experts working in public agencies such as ECHA and follow other actors, namely specialists in materials science, toxicology and computer science as they attempt to craft QSAR models.

Unstable categories, unstable models

QSAR models are based on statistical correlations between a set of physicochemical descriptors which characterize a substance (e.g. chemical composition, morphology…) and its biological activity, which includes its potential toxicity. In other words, QSAR models are based on 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.

One of the main interests of QSAR models for regulatory purposes lies in their ability to re-group chemicals across existing categories and according to similar structure-activity profile. Instead of the existing classifications (such as those based on substances' atomic compositions), substances would be grouped according to their hazard profile. A telling illustration of this point is the case of nanomaterials.

Attempts at regulating nanomaterials within the European institutions have been caught in a tension between two opposite approaches (Laurent, 2017). On the one hand, the European Commission argues for a case-by-case approach to deal with nanomaterials. In this approach, nanomaterials could be gathered in broad categories (carbon nanotubes, titanium dioxide, etc.) each of them further broken down into smaller ones (e.g. single-walled and double-walled nanotubes, rigid single-walled and flexible sing-walled nanotubes…). On the other hand, other regulatory actors criticize this approach for failing to stabilize categories necessary for constraining legal interventions, such as labelling or control. The European Parliament added an amendment to the 2011 cosmetic regulation which introduced mandatory labelling of cosmetics containing nanomaterials. In 2012, France became the first country to introduce a mandatory declaration of nanomaterials. These initiatives require that new definitions be introduced in regulatory texts. In these regulatory texts, nanomaterials were defined using a size limit (set between 1 and 100nm), which could only partly account for the possibility of additional hazard. The antagonism between the two approaches can be summed up as follows: while the former tends to propose an endless subdivision of ever more refined categories (at the price of the postponement of regulatory decision), the latter is based on the construction of general categories, technically imperfect, and possibly arbitrary.

QSAR models can be seen as a way of escaping this quandary. Scientists propose to use QSAR model for nanomaterials, so as to group them in relation to the similarity of different substances' risk profiles. By defining "profiles" of risk more precisely, it would become possible to generate new categories. One could group together substances based on physical or chemical descriptors (e.g. their shapes), and associated expected properties (including those linked with toxicity). Accordingly, QSAR methods would provide a tool to group chemicals according to common characteristics that would generate similar physicochemical properties – including those linked to potential hazards, i.e. the properties that are particularly interesting from a regulatory viewpoint. As such, these methods offer ways of grouping chemicals without either constantly separating them in new categories or creating general and arbitrary criteria.

How is it then possible to group chemicals according to common characteristics correlated with similar properties, including above all toxicological properties? The process we observed when studying scientists developing QSAR for nanomaterials comprised:

- the choice of a set of reference substances (in the project we observed, as many as 45 different nano-substances, belonging to different chemical families such as Zinc oxides, Nickel oxides, or Boehmite);
- the definition of a list of "descriptors" such as the morphology (shape) or the size of chosen compounds, whether they come in filaments, aggregates, etc.;
- the definition of a list of "endpoints" linked to experimental test data on cell cultures in the laboratory, mostly so as to predict rates of reproduction or cell defects;
- the production of statistical correlations between descriptors and endpoints, which led to the refining of both lists.

New groups of chemicals could then be constituted according to their similarities in terms of their structures (descriptors) and correlated activity (endpoints).

The challenge, here, is to avoid two opposite problems. The first one is called *over-fitting* 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 over-fitting, 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 over-fitting, QSAR specialists need to build statistical correlations that are *not too accurate*, in order for the model to be usable for new entry data. Over-fitting 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. *under-fitting*. For a correlation to arise, one needs a minimal number of

descriptors, various enough for statistical relationships to emerge.

Avoiding the problem of over-fitting and that of under-fitting requires that QSAR practitioners proceed with caution. The following discussion (between A, B and C, three members of the research project we observed) is about whether or not to quantify the shape of the substances being used to build the model, and then about what criteria to select in order to differentiate among substances:

A. Descriptors are not all quantitative… how will we do for the shape of substances?
B. 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.
A. But you could also do, "if it's a sphere then 1", "if it's a cylinder 2", "3 is a lump", etc.
B. 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.
C. We could differentiate among 4 types: isotropic isolated nanoparticles, isolated sticks, isolated bars, and formed aggregates. (…)
A. 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.
B. 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.
C. What the chemist sees, and what the biologist sees…
B. But there's no truth in itself here, we choose descriptors from the viewpoint of the cell…
C. 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" from one another, so that a substance on which the model will be used will be described as "particles", "bars", or "aggregates"… Then the question relates to the number and type of these descriptors of shape.

Two remarks follow from there. First, we can see in this exchange that isolating descriptors is a process based on a variety of inputs, including references to guidelines produced by international organizations (here, the OECD), considerations about what will make a difference in toxicological 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 needs 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?).

Therefore, the list of descriptors might significantly vary among QSAR models. In this respect, there is a fundamental uncertainty about the appropriate choice of descriptors, and, consequently, about the categories emerging from the grouping of substances according to descriptors. There is no such thing as "the best" category, but rather a trade-off between different descriptors and the importance granted to various criteria. Getting back to the dialogue above, the project might lead to group substances according to their shapes as "bars" or "sticks", yet will only do so in the context of an inquiry on cell toxicity.

This snapshot is only a glimpse into how QSAR models are produced in practice. One could provide other examples, related not to the choice of descriptors, but also to that of endpoints, or

that of reference substances themselves. Eventually, the calculation of statistical correlations between descriptors and endpoints is itself an iterative process. The person in charge of calculating the statistical correlation between the descriptors and endpoints in the research project that we observed 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. She would eliminate them, thereby reducing an initial long list to just a few parameters.

These considerations show that the practices of QSAR modelling are not stabilized, but partly re-invented for each dataset of chemicals used to build models. For QSAR practitioners, the objective is to build models *accurate enough*. To do so, these practitioners process by trial and error, concerning the list of descriptors, the list of endpoints, and the calculation of statistical correlations. Thus, in QSAR modelling, 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 on / is impacted by the larger choices about their objectives and modes of functioning (MacKenzie, 1993). In this particular case, accuracy is negotiated in a way that never aims to construct the model as a settled entity. Models need to be accurate, yet not too accurate.

This characteristic might result from a more general feature of models based on the identification of statistical correlations, as opposed to models based on the application of general laws of physics or chemistry. Yet in the case of QSAR, they point to particular regulatory issues. This helps to explain the connection between the use of QSAR and considerations related to the 'regulatory context' that was drawn by European actors commenting on the use of this method. Constituting groups of chemicals with similar risk profiles depends on the choice of descriptors and endpoints, the latter being directly tied to regulatory priorities (e.g. aquatic toxicity for certain animal species). Eventually, various choices of descriptors and endpoints might lead to the crafting of various groups of chemicals, each of them tied to certain models. The possibility of re-defining the perimeters of the categories that bring chemicals together is precisely what makes QSAR models interesting in cases such as nanomaterials where substances are not covered by existing regulatory categories. But this also means that QSAR models and the group of chemicals on which they are expected to be applied are constituted in the same movement, and that, consequently, the former cannot easily be disentangled from the latter.

A procedural standardization?

The standardization of models expected to be used for regulatory purposes is a daunting task. Standards for experimental test methods are developed at the Organisation for Economic Cooperation and Development (OECD) and used in the European regulatory bodies. But QSAR models raise practical difficulties for standardization. How, for example, to define in advance the list of descriptors and endpoints without compromising the trial-and-error process that is at the heart of the construction of QSAR models? We begin here to understand the difficulty with which we started our exploration of QSAR models in European regulatory bodies. If the European Chemical Agency does not envision "a formal adoption process, in the same way that test methods are currently adopted in the EU and OECD" (ECHA, 2008: 27, see above), it might well be because of the situatedness of the elaboration of QSAR models.

Yet the regulation of technological innovation provides numerous examples of standardization and/or regulatory interventions that are designed for their ability to cope with the local adaptation of technical tools. Commenting on such processes, Cambrosio and Keating (2009) speak of 'regulatory objectivity'. By contrast with 'mechanical objectivity' (Daston and Galison, 2007), based on stable technical instruments, 'regulatory objectivity' refers to situations within which public and private institutions need to agree on procedures according to which various regulatory entities can be crafted. Regulatory objectivity "consistently results in the production

of conventions, sometimes tacit and unintentional but most often arrived at through concerted programs of collective action" (Cambrosio et al., 2006: 190). Describing various standardization and/or regulatory interventions related to biomedicine, Cambrosio and Keating analyse the ways in which public and private actors coordinate in order to produce procedural instruments ('conventions' or 'protocols') allowing them to stabilize the use of technological tools that might otherwise vary across the local sites where they are applied. Cambrosio and Keating point to a configuration whereby expert judgment may rely on stable tools: where there is no technical black-boxes (e.g. a testing method), then at least a set of agreed principles offers common references for experts to base their actions on. Thus, even if the diversity of QSAR models prevents them from being used as stable instruments that would ensure the production of mechanical objectivity, a procedural approach could be seen as an answer. Since the expert judgment about the hazards of a substance implies a judgment about the validity of the QSAR model being used, then standardized procedures for crafting valid models could be valuable resources. Would an approach based on the standardization of procedures offer a path forwards for experts working in public agencies to use QSAR models?

This directly echoes some of the propositions made at the OECD, where the significant variation of QSAR uses across countries was tied to an issue of harmonization:

> 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)

In this quote, "the development of a set of agreed principles" can be read in the terms of regulatory objectivity. It proposes international coordination for producing conventions. Within the international organization, this objective is directly connected to a boundary work, between internationally harmonized procedures that could guarantee the validity of the modelling approach, and the technical content of the model, which could be adapted to local situations according to regulatory choices (Thoreau, 2016). The task of the international organization, here, is to define generic principles of use, defined in such ways that they do not cross the perimeter of states' regulatory choices. Distinguishing international principles from (nationally-produced) technical content is both a way of standardizing QSAR models through conventions and ensuring international agreement without delving into potentially contentious regulatory choices.

The principles that the OECD released were 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 predictability;
> 5. a mechanistic interpretation, if possible (OECD, 2007: 14).

These guidelines offered a way of ensuring international agreement about QSAR validation processes. Yet these principles had to do so without entering the domain of regulation, which is that of sovereign policy choices, and outside the scope of OECD intervention. Thus, instead of stating which endpoints or which algorithms should be used (choices potentially related to regulatory decisions), the guidelines stated that the two had to be identified in unambiguous ways. For the OECD intervention to be acceptable, QSAR validation principles had to be framed in a very general way.

The attempt to craft principles according to which the quality of QSAR models could be assessed is directly connected to a crucial issue for model-making, namely validation. Validating a model is both a technical task, tied to the scientific value of the model, and a political one, as it must be decided whether or not the model is robust enough to ground policy action (Edwards, 1999). While the OECD principles only considered the validation of QSAR models in general terms so that the international organization would not enter the perimeter of states' regulatory actions, the European institutions undertook an explicit reflection about whether and how QSAR models could be validated.

Validating QSAR models can be carried out by 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.

Considering the diversity and permanent evolution of statistical tools, Andrew Worth, QSAR specialist and Senior Scientific Officer at the Joint Research Centre (JRC) of the European Commission, concludes that the validity of a given model 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-examine the appropriate validation methods for each new situation.

Situated expert judgment and the QSAR toolbox

Validation processes can only take the form of general prescriptions. This makes it impossible to consider QSAR models as stable black-boxes that could circulate straightforwardly across various domains of application. This does not mean that standardization is impossible, but that this standardization cannot take the form of technical harmonization (if, for instance, descriptors or endpoints were predefined) or procedural harmonization (if widely applicable validation principles were identified). Both types of harmonization (technical and procedural) require a certain stability of the technology being standardized, whether a stable instrument turned into a black-box circulates across various sites of application, or stable principles define procedures expected to be generally applicable. This means that QSAR models cannot be grounded on mechanical objectivity and the accompanying 'view from nowhere', or on regulatory objectivity and the coordinated approach on which it relies. How then can we understand the type of expert judgment at play when QSAR models are used? Another OECD initiative, the "QSAR toolbox" developed in partnership with the European Chemical Agency (ECHA), can help us to understand how experts working in public agencies are expected to use QSAR models.

Developed at the OECD and supported by ECHA since 2008, the QSAR toolbox is a free software application designed to "identify and fill (eco)toxicological data gaps for chemicals hazard assessment" (ECHA, 2011). It is intended to be used by private companies seeking to evaluate the hazard of the substances they produce, by experts working in public agencies and in charge of eval-

uating companies' propositions, and by other stakeholders[2]. Contrary to what its name seems to indicate, the QSAR toolbox does not provide a ready-made QSAR model fit for application on any given chemical. Rather, it brings together:

- databases with results from experimental studies;
- accumulated knowledge for structural characteristics (alerts) that can indicate the presence of hazards and other properties, and
- tools to estimate missing experimental values by read-across, by trend analysis (i.e. interpolating [preferred] or extrapolating from a trend [increasing, decreasing, or constant] from tested to untested chemicals within a category) and/or by (Q)SAR models. (ECHA, 2011)

Thus, QSAR models are one component of a more general platform. This platform is fed with experimental data, some of which are related to the physical causality between "structural characteristics" and hazards (second bullet point in the previous quote), and comprise modelling tools, some quantitative (as QSAR models are), and others based on statistical approaches that do not use quantitative predictive modelling. An example of the latter in the quote above is "read-across", which consists in using available empirical data to estimate the missing ones. The QSAR toolbox does not attempt to deliver ready-made risk assessments for a user (whether a regulator or a scientist) eager to know the toxicity of a given chemical. Rather, it offers a way "to systematically group chemicals into categories according to the presence or potency of a particular effect for all members of the category." (ECHA, 2011). "A particular effect" relates here to the particular endpoint that the user might want to test, and which requires the mobilization of various experimental data and instruments, comprising QSAR models and other, non-quantified, statistical tools.

Rather than providing a neatly defined quantitative instrument to which the technical task of risk assessment could be delegated straightforwardly, the QSAR toolbox is a platform that demands a reflective and cautious intervention by users, as they work on its many components to gather a set of indications about whether a chemical could be grouped with others, and how so. The OECD (2007: 92) gives the example of choosing a "no-observed-effect" as an endpoint. It asserts that while such a level may be relevant for policy-making purposes, it may as well be irrelevant for the purpose of generating scientific knowledge, i.e. "referring to a specific effect within a specific tissue/organ under specified conditions" (OECD, 2007: 92). One sees here that an active uptake about the very purpose of choosing the endpoint will affect its relevance.

Thus, the QSAR toolbox can only be used by an informed user, who has particular regulatory objectives in mind. This informed user is able to identify the scope of the evidence provided, and its limitations. This means that the QSAR toolbox can in no way be mobilized as a black-boxed instrument that could be used without opening up its inner mechanism. It follows that the concern for the transparency of the platform is constant among both the designers and users of the QSAR toolbox. Allowing regulators to access the characteristics of databases has become a necessary condition for the platform to function, as an OECD official told us during an interview:

> 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. (...) 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)

Being transparent about the toolbox is about making its inner mechanism visible. It is also about making it possible for users to contribute, by providing new experimental data that could refine the existing correlations. The toolbox is indeed designed to be fed on an on-going basis with new experimental data and refined statistical correlations. Such a development implies enrolling more and more users, so as to ensure both the collective legitimacy and the technical validity of the instrument. This enrolment process is driven by the constitutive process of the toolbox itself as depicted

above. It follows that it cannot be considered as a mere "beta testing" phase after which the toolbox would be closed and remained unchanged. Instead, openness, try-outs and transparency are inherent to the exercise of QSAR modelling.

The case of the QSAR toolbox is particularly interesting to further our understanding of the difficulty related to the use of QSAR models. Many ECHA documents state the impossibility of envisioning a formal adoption process of QSAR models within the European regulation of chemicals (see section 1). It is a consequence of the approach lying at the heart of the QSAR approach, and, eventually, a consequence of the particular type of standardization that can be pursued. Rather than standardizing a technical content or a procedure, the OECD and ECHA proposed a constantly evolving platform expected to help its users group chemicals together, along lines that are permanently subject to change.

The QSAR toolbox is meant to make QSAR models usable. Examining how it does so, as we have just done, is a way of better identifying the characteristics of the QSAR models, and the ways in which they are expected to contribute to risk governance within the QSAR toolbox:

- QSAR models are constituted at the same time as the groups of chemicals which they are expected to govern, and cannot easily be disentangled from these groups;
- Their scientific and regulatory value can only be assessed according to general criteria, which then require case-by-case assessment of models;
- QSAR models are not stable entities circulating across situations of use. Rather, they are meant to be articulated with one another and with other methods (as in the QSAR toolbox), so as to be refined as new experimental data are produced;
- Therefore, their potential users are not expected to apply them as ready-made instruments that operate autonomously, but need to mobilize their informed judgment to assess the ways in which they can provide relevant information for a given regulatory purpose. This results in an emphasis on transparency.

All these characteristics made QSAR models unfit for standardization as black-boxed instruments. Private companies and public experts can use them in coordination with other approaches. A platform such as the QSAR toolbox is therefore better defined as a 'grey box', which is mobilized in different ways according to particular situations of use, and never meant to be closed to external examination. Eventually, the QSAR toolbox cannot serve as an unproblematic coordination device, which could guarantee the value of the risk assessment performed by private companies and could be used by public experts to validate it. The toolbox example provides an illustration of how expert judgment is expected to be exercised in the case of QSAR. Rather than grounding the expert intervention on the ability to mobilize stable instruments that make the individual characteristics of the expert disappear (as when mechanical objectivity is the objective) or on the possibility to refer to common procedures (as in a regulatory objectivity framework), QSAR models require expert judgement to be situated locally, and discussed in relation with particular regulatory objectives. This has consequences for risk governance, which the next and last section discusses.

What risk governance in the world of QSAR models?

So far, we have discussed how QSAR specialists craft their models, how the OECD proposes only general principles of validation and a QSAR toolbox that is neither the provider of ready-made instruments nor the vehicle for common and operational procedures. What about the work of people in charge of evaluating the proposals of companies attempting to register the substances they produce? This is the task of public experts working at the European Chemicals Agency, and at national agencies in charge of risk assessment. Our reflection started with the consideration of the practical difficulties that these actors encountered when using QSAR models. These experts working in public agencies do not develop QSAR models. Nor are they in charge of standardizing their use[3]. Instead, they need to evaluate the ways in which companies describe the risk profile of the

substances they wish to register. The impossibility of using QSAR models as black boxes, and the mobilization of grey boxes such as the QSAR toolbox, has consequences on how they can assess the validity of companies' claims.

First, public agencies constantly need to examine the QSAR models used by companies. Consider for instance how members of the French public agency for environmental safety describe their roles in assessing how companies use QSAR models:

> - And I think that the challenge for us is to identify the limits and confront the companies. (…) 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)

This quote points to an important consequence of the use of QSAR models for risk assessment purposes. Because of the complexity of these methods, and the diversity of actors producing them (in various ways according to the particularities of the situation), public experts might find themselves in a position of weakness - as they need to assess pieces of evidence produced by non-standardized and ever more complex tools. This asymmetry is only made more acute by the diversity of actors producing QSAR models. In addition to public research centres, many companies and open-source communities also develop their own QSAR software, either licensed or not, for profit or not (Lo Piparo and Worth, 2010). Datasets to inform the models are compiled by many different actors, including scientists for knowledge-production purposes, but not only. Many statistical techniques or mathematical models can be tailored to the creation of a particular QSAR. Various heuristic tools and different classes of algorithms are designed as a means to browse through the diversity of data and gather different sorts of results, including a wealth of machine-learning techniques (Lavecchia, 2015).

Second, the nature of expert intervention evolves, as neither the delegation to a trusted instrument (as in a regime characterized by mechanical objectivity) nor the mobilization of collectively produced conventions (as in a regime of regulatory objectivity) are possible. When assessing the dossiers submitted by companies to apply for the registration of chemicals, officials at ECHA will examine the models by opening them up, and comparing them with experimental data, as one of them told us during an interview:

> If I 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. (Interview, ECHA)

This quote explicitly connects the diversity of the methods used to produce evidence that require transparency (the expert needs to know what is inside the models) with the possibility for the expert working in public agencies to draw on other sources of information. The same official eventually referred to experts' "own experience" in assessing the use of QSAR models:

> 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, emphasis added)[4]

When confronted with QSAR models, expert judgment is based on the expert's experience, and on his ability to confront the construction of the model itself with other sources of information. This directly echoes the expected functioning of the QSAR toolbox (see above). Yet it stands in uneasy relation to the complexity of QSAR models, as the potential sophistication of the statistical

approaches might well turn some QSAR models into black-boxes that are impossible to open to the gaze of the experienced public expert.

A condition for carrying out such a situated expert judgment is that public experts have the possibility to access the inner functioning of the models presented to them. During an interview, an ECHA official explained the issue this situation raised in the following terms:

> 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

Thus, the requirement for transparency makes public experts wary of overly complex instruments that they would be unable to grasp (regarding, for example, the hypothesis, the domain of applicability, or the statistical methods being used). But the requirement for transparency also impacts the institutional role of public expert bodies. We showed in that the ways in which QSAR models are constructed is directly tied to regulatory objectives (as, for instance, the set of endpoints is chosen according to regulatory requirements, or the models and the group of chemicals on which they are expected to be applied are manufactured in the same process). This means that when examining companies' use of QSAR models, public experts working in agencies such as the ECHA also need to evaluate how model-based risk assessment approaches fit with regulatory objectives.

We can now get back to the three-body problem of expertise. The examination of the practical conduct of QSAR modelling and standardization shows that QSAR models challenge the three components of expert legitimacy. Rather than grounding expert judgment in the ability to deliver a 'view from nowhere' in which the individuality of the expert disappears, QSAR makes public experts fully- fledged individuals who need to draw on their personal experience to evaluate private actors' propositions. Rather than providing a set body of knowledge, possibly formalized in black-boxed instruments or standardized by stable procedures, the use of QSAR for the regulation of chemicals requires situated examinations that need to be adapted to the particularities of every case. Rather than being a public body intended to act in isolation as a provider of scientific advice to inform risk management decisions, from which it is neatly separated, ECHA is an institution that coordinates with national agencies and international organisations in developing tools for the evaluation of QSAR models, while articulating its risk assessment mission with regulatory considerations.

Conclusion

As models are increasingly expected to contribute to regulatory decisions, understanding their political consequences it crucial. This paper has focused on models aiming to produce statistical correlation, and discussed one of these political consequences, related to the type of expert judgement that models entail. The case of QSAR models in the European governance of chemical risks illustrates a type of expert judgment that is situated, as experts in regulatory bodies cannot consider models as stable technical black boxes, and cannot rely on standardized procedure to use them. This explains why QSAR models, while being seen as a powerful alternative to animal testing, are also considered with caution in expert bodies such as the European Chemicals Agency. Examining how QSAR models are crafted in practice, we showed that this situation is not the first step before these models can act as stable instruments, but is derived from their very characteristics (and of what makes them interesting in the first place). Attempts are made to standardize principles for their evaluation, and a "toolbox" is proposed by the OECD to carry forth their validation. Yet, the use of QSAR models does not imply either mechanical objectivity or regulatory objec-

tivity. Instead, the use of QSAR models in the European regulatory context means that expert judgment is situated, and grounded in the experience of the expert.

Facing the proliferation of information produced by models that they cannot completely rely on, public experts are confronted with an asymmetry of resources. They need to invent procedures by which they can gather enough information to make regulatory choices. For private companies, the submission of dossiers is becoming more strategic than ever, since the plurality of available models means that some of them might suit their needs and interests better than others. These companies therefore mobilize resources and develop an in-house expertise on models in their routine R&D process. This results in increasing demands on public bodies in charge of critically examining the models used by industries. That transparency becomes a growing concern follows, since public experts need to open up the models, or at least gather information about them. As 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 experts need to re-invent their roles so that they are able to monitor both the construction and the use of models.

The case of models in the governance of chemicals is specific, yet has value for a broader reflection on the use of models for regulatory purposes, particularly correlation-based statistical instruments, as QSAR models are. The value of these particular models is tied to the empirical data they are based on, and to the domain of use they are applied to. This paper has shown that the type of objective knowledge that these models are claimed to produce requires an active intervention of the expert in charge of interpreting it. As models are increasingly called for to settle controversies, plan long-term developments, or argue for or against policy choices, it is crucial not to see them as ready-made providers of objective knowledge, but as instruments that re-work what objectivity is, and directly constrain how experts can and should act.

Acknowlegments

A previous version of this paper was discussed during the workshop "Produire la prediction: le travail de la modélisation pour l'évaluation des risques", held in Paris, June 19th, 2017, at the invitation of David Demortain and Henri Boullier (project INNOX). The authors would like to thank warmly the anonymous reviewers, as well as the editors of this special issue, for their insightful comments and suggestions.

References

Boullier H (2016) *Autoriser pour interdire. La fabrique des savoirs sur les molécules et leurs risques dans le règlement européen REACH. Thèse pour le doctorat de sociologie*. Université Paris-Est Marne-la-Vallée.

Boullier H, Demortain D and Zeeman M (2019) Inventing Prediction for Regulation: Modelling Structure-Activity Relationships at the US Environmental Protection Agency. *Science & Technology Studies* 34(2): 137-157.

Cambrosio A and Keating P (2009) Biomedical Conventions and Regulatory Objectivity. A Few Introductory Remarks. *Social Studies of Science* 39(5): 651-664.

Cambrosio A, Keating P, Schlich T and Weisz G (2006) Regulatory objectivity and the generation and management of evidence in medicine. *Social Science & Medicine* 63: 189–199.

Daston L and Galison P (2007) *Objectivity*. New York: Zone books.

Demortain D (2009) Legitimation by standards: Transnational experts, the European Commission and regulation of novel foods. *Sociologie du Travail* 51:2 104-116.

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

ECHA (2011) The OECD QSAR toolbox for grouping chemicals into categories. *European Chemical Agency*. ECHA-11-L-08-EN.

Edwards P N (1999) Global climate science, uncertainty and politics: Data-laden models, model-filtered data. *Science as Culture* 8(4): 437-472.

Edwards P N (2010) *A vast machine: Computer models, climate data, and the politics of global warming*. Cambridge (MA): MIT Press.

European Commission (2006) *Regulation (EC) n° 1907/2006 of 18 December 2006 Concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH)*. Consolidated version of 1 June 2015, doc. 2006R1907. Luxembourg: Publications Office of the European Union.

European Commission (2016) Better Regulation: Delivering better results for a stronger Union. *Communication from the Commission to the European Parliament, the European Council and the Council*. COM (2016) 615 final.

Fisher E, Pascual P and Wagner W (2010). Understanding environmental models in their legal and regulatory context. *Journal of Environmental Law* 22(2): 251–283.

Gramatica P (2007) Principles of QSAR models validation: internal and external. *Molecular informatics*, 26(5): 694-701.

Heaphy L (2015) The role of climate models in adaptation decision-making: the case of the UK climate projections 2009. *European Journal for Philosophy of Science* 5(2): 233-257.

Hilgartner S (2002) *Science on Stage. Expert Advice as Public Drama*. Stanford: Stanford University Press.

Jasanoff S (1990) *The Fifth Branch. Science Advisers as Policymakers.* Cambridge (MA): Harvard University Press.

Jasanoff S (2004) *States of Knowledge*. Cambridge (MA): MIT Press.

Jasanoff S (2005) Judgment Under Siege: The Three-Body Problem of Expert Legitimacy. In: Maasen S and Weingart P (eds) *Democratization of Expertise? Sociology of the Sciences Yearbook*, vol 24. Dordrecht: Springer, pp. 209-224. .

Jasanoff S (2011) The practices of objectivity in regulatory science. In: Camic C, Gross N and Lamont M (eds) *Social Knowledge in the Making*. Chicago: University of Chicago Press, pp. 307-337.

Latour B (1987) *Science in action: How to follow scientists and engineers through society*. Cambridge (MA): Harvard university press.

Laurent B (2017) *Democratic experiments. Problematizing nanotechnology and democracy in Europe and the United States*. Cambridge (MA): MIT Press.

Lavecchia A (2015) Machine-learning approaches in drug discovery: methods and applications. *Drug Discovery Today* 20(3): 318–31.

Lo Piparo E and Worth A (2010) Review of QSAR Models and Software Tools for predicting Developmental and Reproductive Toxicity. *JRC scientific and technical reports* EUR, 24522.

MacKenzie D (1993) *Inventing Accuracy. A Historical Sociology of Nuclear Missile Guidance*. Cambridge (MA): MIT Press.

OECD (2007) *Guidance document on the validation of (quantitative) structure-activity relationships [(Q)SAR] models*. OECD, 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.

Saurugger S (2002) L'expertise: un mode de participation des groupes d'intérêt au processus décisionnel communautaire. *Revue française de science politique*, 4:52, 375-401.

Thoreau F (2016) 'A mechanistic interpretation, if possible': How does predictive modelling causality affect the regulation of chemicals? *Big Data & Society* July-December: 1-11.

Vos E (2000) EU food safety regulation in the aftermath of the BSE crisis. *Journal of Consumer Policy* 23(3): 227-255.

Wickson F and Wynne B (2012) The anglerfish deception. The light of proposed reform in the regulation of GM crops hides underlying problems in EU science and governance. *EMBO reports* 13: 100-105.

Worth A, Hartung T and Van Leeuwen CJ (2004) The role of ECVAM in the validation of QSARs. *SAR and QSAR in Environmental Research* 15(5–6): 345–358.

Worth A, Bassan A, De Brujin J et al. (2007) The role of the European Chemicals Bureau in promoting the regulatory use of (Q)SAR methods. *SAR and QSAR in Environmental Research* 18(1-2): 111-125.

Notes

1 We were co-partners of this research project, in charge of analysing the political aspects of models meant for regulatory purposes.

2 See OECD Toolbox website: https://www.qsartoolbox.org (accessed 2017-12-05).

3 The distinction is partly arbitrary since the ECHA experts also participate in the OECD working groups.

4 In this quote, 'regulators' is used to qualify the experts working in public agencies.