Information, Communication & Society

ISSN: 1369-118X (Print) 1468-4462 (Online) Journal homepage: https://www.tandfonline.com/loi/rics20

The (in)credibility of algorithmic models to non-experts

Daan Kolkman

To cite this article: Daan Kolkman (2020): The (in)credibility of algorithmic models to non-experts, Information, Communication & Society, DOI: 10.1080/1369118X.2020.1761860

To link to this article: https://doi.org/10.1080/1369118X.2020.1761860

Published online: 18 May 2020.

INFORMATION, COMMUNICATION & SOCIETY
https://doi.org/10.1080/1369118X.2020.1761860

Routledge
Taylor & Francis Group

OPEN ACCESS

The (in)credibility of algorithmic models to non-experts

Daan Kolkman [a,b]

[a]Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands; [b]Jheronimus Academy of Data Science, 's-Hertogenbosch, The Netherlands

ABSTRACT
The rapid development and dissemination of data analysis techniques permits the creation of ever more intricate algorithmic models. Such models are simultaneously the vehicle and outcome of quantification practices and embody a worldview with associated norms and values. A set of specialist skills is required to create, use, or interpret algorithmic models. The mechanics of an algorithmic model may be hard to comprehend for experts and can be virtually incomprehensible to non-experts. This is of consequence because such black boxing can introduce power asymmetries and may obscure bias. This paper explores the practices through which experts and non-experts determine the credibility of algorithmic models. It concludes that (1) transparency to (non-)experts is at best problematic and at worst unattainable; (2) authoritative models may come to dictate what types of policies are considered feasible; (3) several of the advantages attributed to the use of quantifications do not hold in policy making contexts.

ARTICLE HISTORY
Received 18 December 2019
Accepted 16 April 2020

KEYWORDS
Algorithms; ethnography; credibility; quantification; decision making

Introduction

The ongoing increase of computational power allows for the development of ever more sophisticated data analysis techniques, methods, and algorithms (Venturini et al., 2015). Algorithms in particular are imbued with promises of reliability and objectivity (Mazzotti, 2017) and are considered instrumental towards solving pressing societal issues such as climate change and mass-migration (Floridi & Taddeo, 2016). The perceived usefulness of algorithms extends beyond their superior accuracy and includes communicative and organizational benefits (Bissell et al., 2012; van Daalen et al., 2002) which can be grouped under the concept of commensurability (Kuhn, 1962).

At the same time, the increasing influence of algorithms has been cause for concern (Dourish, 2016) pertaining loss of privacy (Taylor, 2017), bias (Barocas & Selbst, 2016; Boyd & Crawford, 2012), and racism (van Doorn, 2017). A growing number of scholars conduct social studies of algorithms and interrogate their impact from an ethical or legal perspective yet understanding of how people 'do' data analysis and use algorithms has lagged behind (Christin, 2017). In light of recent attempts to regulate algorithms

CONTACT Daan Kolkman d.kolkman@tue.nl Sint-Janssingel 92, 5211 DA 's-Hertogenbosch, The Netherlands
This article has been republished with minor changes. These changes do not impact the academic content of the article.

(see Goodman & Flaxman, 2017) and many examples of algorithms gone awry (e.g., Barocas & Selbst, 2016; Pielke, 1999; Van der Sluijs, 2002) such research is long overdue. This is particularly pressing considering modelling underpins government decision making in relation to the COVID-19 crisis.

This paper answers Christin's (2017) call for more ethnographic work that considers the use of algorithms in practice. More specifically, I problematize the dichotomy between the perceived benefits and observed shortcomings of algorithms and consider how those working with algorithms assess the credibility of these algorithms. I focus on a specific subset of algorithms used in the context of policy making.[1] Kemper and Kolkman (2019, p. 1) refer to these algorithms as *algorithmic models*: 'formal representation[s] of an object that an observer can use to answer questions about that object'. Following Cardon et al. (2018), the algorithmic models studied in this paper resemble hypothetical-deductive machines rather than inductive machines.[2] That is, the starting point of their developers was a conceptual model of the object of interest, which was then parametrized using some optimization algorithm. None of the algorithmic models discussed here was developed following a pure machine learning approach where feature engineering, feature selection, and model selection were fully delegated to the machine. I consider these models[3] and the contexts of their use as *socio-technical assemblages* (Latour, 1987) and draw on eight case studies conducted in government in the form of interviews, participant observations, and documents analysis.

The paper is structured as follows. First, I shortly discuss my data, methods, and the concept of credibility. Subsequently, I draw on empirical material to argue that models are codified quantifications which require a set of specialist skills to operate and show that models are adjusted over time which further strains their scrutability. I proceed by detailing four sets of practices through which those working with models establish and ascertain their credibility. This informs a discussion of the power asymmetries which may emerge as a consequence of the time, skills, and resources required to engage in the different sets of practices. I conclude that: (1) transparency to (non-)experts is at best problematic and at worst unattainable; (2) authoritative models may come to dictate what types of policies are considered feasible; (3) several of the advantages attributed to the use of quantifications do no hold in policy making contexts.

Data and methods

I entered the field as a data science practitioner and sociologist. In response to an overwhelming body of technical research and pervasive quality assurance guidelines, I sought to describe the social context in which models are used and provide insights into what makes models credible to those who use them. It is important to mention that my informants used many concepts to describe their work, some identified as 'policy analysts', others as 'modellers', and others still as 'data scientists'. I use the term model-professional to refer to this loose association of people whose work in one way or another involves a model. The term 'model-professional' reflects my intention not to differentiate between my informants in terms of their competence; all were taken to be skilled in the interactions with the model which they engage in routinely. I use the term expert to refer to those model-professionals with a competence in the technical aspects of modelling. In my analysis, I aimed to go beyond descriptions of narratives and aimed to include the day-to-day work involved in the development, use, and interpretation of quantifications. I studied eight cases of model use in government over a period of 2.5 years. Between 2013 and 2016, data were collected on these eight models in the form of interviews, documents and observations. The data included 38 semi-structured interviews with model developers and policy analysts. The interview data were supplemented by an archival study of 41 documents such as minutes, model documentation and policy documents that reference the models. In addition, in situ observations of model developers and policy analysts were conducted. I have presented a step-by-step overview of the sampling methodology, a detailed description of the case studies, and a quantitative analysis elsewhere (Kolkman, 2020; Kolkman et al., 2016). Here I repeat the most important considerations of the data collection and subsequent analysis. The cases were sampled following a maximum variation regime (Patton, 2002) from publicly available inventories of models, I employed a set of selection criteria that were suggested by previous work on model use: Novelty (Venkatesh & Bala, 2008), model technique (Boulanger & Bréchet, 2005); model purpose (Treasury, 2013), and cultural differences (Diez & McIntosh, 2011). An overview of the cases I selected can be found in Table 1.

Following Heuts and Mol's (2013), I invited my informants to talk about their day-to-day activities and asked them to explain the aspects of their work that they might take for granted. I iteratively coded and rearranged it to foreground the practices that model-professionals engage in. I employed a format based on the grounded theory paradigm put forward by Strauss and Corbin (1998). All fragments of text in each interview, document, and observation note were open coded. Open codes were redefined and recombined to develop clearer concepts per case study. An overview of the resulting codes was written up and presented to my informants for validation. I rearranged my material by codes and extracted those codes that related to credibility practices. To protect my informants, I have removed identifying information where possible and used pseudonyms to refer to them.

From the outset of the coding process, I defined credibility as the quality of a particular piece of evidence, since this is how it has typically been defined in the context of policy making (see Cash & Clark, 2001). However, the fundamental question that is of interest is how and why people come to trust a particular piece of evidence (Head, 2010). Consider the following excerpt from an interview with one of my informants:

> George: On the whole, everyone sort of suspends their disbelief to some extent around how precise Pensim2 is. It allows the process to then happen in a much better way. Because everyone sort of agrees: 'We want it to be kind of cost neutral and we'll take the model on trust and if we come up with something and the model says it's cost neutral then it's not too far off'. And everyone can be happy that the costs are fairly robust, and the negotiations can happen of one agreed view of the world.

As such, I moved away from considering credibility as a fixed characteristic of a particular piece of evidence – in this case a model – and gravitated towards approaching it as something that is *enacted* (Law & Mol, 2006). Credibility can change over time, needs to be developed, and maintained by the people who work with a model. My focus is thus on the practices through which 'everyone sort of suspends their disbelief' and agrees that a model is credible, despite possible shortcomings. Here, I take 'everyone' to mean the model-professionals who work with a particular model, as conceptions of credibility

Table 1. An overview of the case studies and their characteristics.

	Model characteristics		Context		
Case ID	*Novelty*	*Technique*	*Purpose*	*Location*	*Organisation*
Pensim 2	>10 years	Dynamic microsimulation	Policy simulation	United Kingdom	Department for Work and Pensions
SAFFIER II	>15 years	Macro-economic model	Forecasting	The Netherlands	Netherlands Bureau for Economic Policy Analysis
WEF	>5 years	Linear regression	Planning	The Netherlands	Netherlands Land Registry
EUPA	>5 years	Accounting model	Policy simulation	United Kingdom	Operated by a consultancy for the Department for Environment, Food and Rural Affairs
2050 Calculator	<2 years	Accounting model	Science-based model	United Kingdom	Department for Energy and Climate Change
RRI	<2 years	Decision tree model	Financial evaluation	The Netherlands	Operated by a consultancy for a variety of clients
UKTimes	<1 year	Linear optimization model	Forecasting	United Kingdom	Department for Energy and Climate Change
Quitsim	<5 years	Agent-based model	Policy simulation	United Kingdom	Operated by a consultancy for Public Health England

may vary from one community to the next (Shapin, 1995). This variability is paramount in understanding quantification objects (Espeland & Stevens, 2008) and sociologically important, both in terms of discourse (Gilbert & Mulkay, 1984) and practices (Heuts & Mol, 2013). Although variability can be considered a key component to the analysis, this does not imply an absence of structure. Gilbert and Mulkay (1984) and Heuts and Mol (2013) identify such arrangements in their data and refer to them as ‘repertoires’ and ‘registers’, respectively. The point of such overarching arrangements is not to provide a ‘x’ number of dimensions that constitute a phenomenon; the dimensions are neither mutually exclusive nor collectively exhaustive. Rather, these arrangements serve as an analytic lens to uncover salient characteristics of the socio-technical assemblage. In my analysis I distinguish between two types of such arrangements: sets of practices and repertoires. The former refers to collections of practices that are similar, the latter points towards the different availability of practices to those who work with models.

Analysis: four sets of credibility practices

General observations

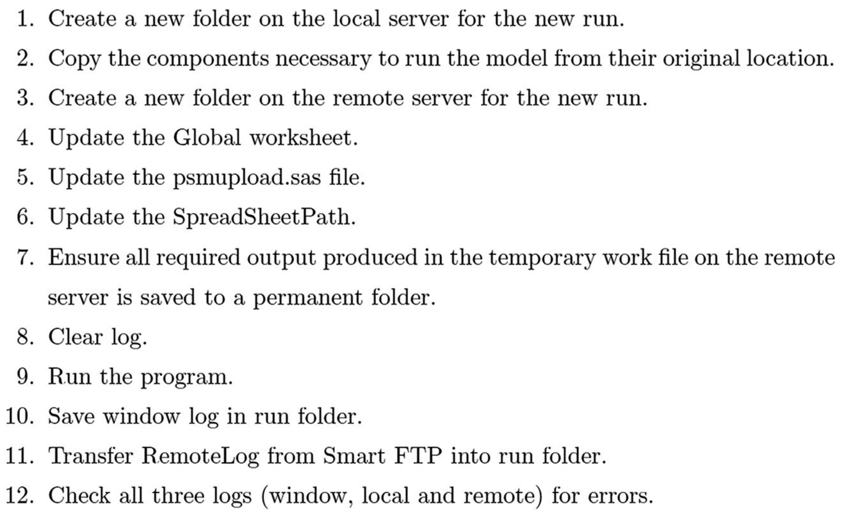
In what follows, I discuss two general observations that are of consequence to model use in practice. First, I argue that models are codified quantifications which require a set of specific skills. Secondly, I show that models are dynamic and are adjusted over time which results in documentation and explanations lagging behind. I proceed by introducing and discussing the four sets of credibility practices that model-professionals engage in: intuitive – formal – procedural-, and evaluative practices. Finally, I explore how differences in the repertoires introduce and maintain power asymmetries between model-professionals. Quantifications like models require a specific set of skills and a certain level of technical knowledge to develop or operate. Some of the model-professionals I talked to did have this expertise, while others did not. My informants referred to operating a model as ‘running’ a model. Typically, such a run of the model consists of a series of steps and may include: the selection of input data, configuration of model parameters, initiating the model, and interpreting the model outcomes. In some cases, informants followed a predefined series of steps, while in other cases the process was less structured. Figure 1 provides an example of a more structured approach to running a model. Such a series of steps may look straightforward, yet model-professionals report that it can take considerable time to develop the confidence to run a model without assistance. Closer inspection of the steps reveals why. In step four Pensim2, for instance, model-professionals have to 'Update the Global worksheet'. As can be seen in Figure 2, this worksheet contains thirteen settings which the user has to specify. The configuration options include abbreviations and jargon that novice Pensim2 professionals may be unfamiliar with.

This type of *codification* was not exceptional to Pensim2; it was a feature of all the models I studied. In some cases, model-professionals reported that it took up to a year before they attained the level of skill – and confidence – necessary to run a model independently. Even for expert model-professionals with experience in using other models, learning the specific terminology of a new model takes time. One of my informants reflects on this learning process and explains:

> George: In the end it is quite a lot of detective work in trying to learn it, because of the lots of variables relating to particular components of the pension calculation. It's very easy to assume that a variable may mean something else. And there is a list of all the variables in the main global workbook, which is reasonably good. But in the end, there is a fair bit of detective work. If you want to do detailed changes or do a very detailed picture of who gains or who loses from a particular element of your reform.

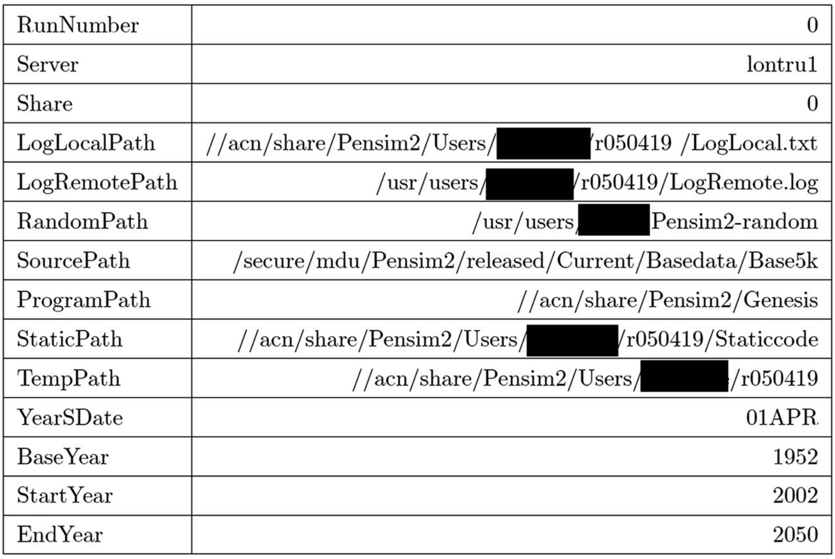
Because of the long period associated with learning how to use a model, training model experts is a costly affair. However, even after a model-professional has completed his or her training, it is not uncommon for them to get lost in the black box of a model. Leo explains:

> Leo: You get to a point where you sort off know it, but you have to be careful. Because sometimes when you're feeling gets too good, it turns out you are being too complacent about something. There is often something horrible and then you find out: "Oh, that's how that works."



1. Create a new folder on the local server for the new run.
2. Copy the components necessary to run the model from their original location.
3. Create a new folder on the remote server for the new run.
4. Update the Global worksheet.
5. Update the psmupload.sas file.
6. Update the SpreadSheetPath.
7. Ensure all required output produced in the temporary work file on the remote server is saved to a permanent folder.
8. Clear log.
9. Run the program.
10. Save window log in run folder.
11. Transfer RemoteLog from Smart FTP into run folder.
12. Check all three logs (window, local and remote) for errors.

Figure 1. Steps for a Pensim2 run (taken form the model documentation).



RunNumber	0
Server	lontru1
Share	0
LogLocalPath	//acn/share/Pensim2/Users/[redacted]r050419 /LogLocal.txt
LogRemotePath	/usr/users/[redacted]/r050419/LogRemote.log
RandomPath	/usr/users[redacted]Pensim2-random
SourcePath	/secure/mdu/Pensim2/released/Current/Basedata/Base5k
ProgramPath	//acn/share/Pensim2/Genesis
StaticPath	//acn/share/Pensim2/Users/[redacted]/r050419/Staticcode
TempPath	//acn/share/Pensim2/Users/[redacted]/r050419
YearSDate	01APR
BaseYear	1952
StartYear	2002
EndYear	2050

Figure 2. Pensim2 Global worksheet configurations.

In the cases I studied, my informants suggested only a handful of people within the organization had the necessary skills and know-how to run the model independently. A much larger group of people worked with the black boxed outputs of the model, for instance, to write policy documentation or brief senior civil servants. Only a few model-professionals had the experience to oversee the model in its entirety, to the extent that they felt they could contribute to its development:

> Sebastian: How the regression equations are derived? I don't know a great deal about how they are derived. I've seen a lot of the spreadsheets, the workbooks and know how they are working, taking whatever variable and applying this parameter, but then I have no idea how those parameters were developed in the first place. I know they are the outcome of some kind of logistic regression, but I don't know much more than that.

When asked about Pensim2, Sebastian pointed out that he did not have complete knowledge of that model even though he was one of its developers. This demonstrates that even those who work with a model closely may not comprehend it in its entirety, yet a model may still be used as a basis for policy making. The terminology of models makes them black boxes that are hard to grasp for expert model-professionals and this codification presents a major hurdle to non-experts. This is exacerbated because models are rarely static. Rather, models are dynamic; they develop over time through a process of continuous adaption. Expert model-professionals need to update the model for it to remain aligned with the latest technical insights, perceptions of the target system, and the policy agenda. A model with the same name may be vastly different today than it was a year ago. Lars explains why such changes are necessary.

> Lars: An example of a study we did for the model pertains to the housing market; what are the effects of a decrease in the value of residential property on consumption? Those can be different in times of sharp economic decline. (…) We don't always get a clear-cut answer, but if we find "something is off in our" model, we will conduct further research into that. You would

> hope to find something you can include in the model. That's the way we approach it. The world itself changes as well. So, you have to re-estimate your model every so often to keep it up to date.

Moreover, multiple versions of a single model may co-exist and may be referred to by same name. This introduces additional difficulty for those looking to understand a model and illustrates why model-professionals may struggle to learn how to run a model. Moreover, it presents an impediment to using model plus judgement decision making approaches in practice. As one of my informants put it:

> Jan: You have to have knowledge of the model. You have to know how it works. And you have to have some experience with its outcomes to be able to appreciate the model. You also need to be aware of the relativity of outcomes. If you are further away from the model, you might think: "the model is the model". However, it is important to adopt a critical attitude. You have to know what the model is, and what it's not.

Several of my expert informants expressed that they felt that there was 'no way' a non-expert model-professional can comprehend the full technical details of a model. This begs the question if and how such non-expert model-professionals pry open the black box and ascertain the credibility of a model. In the following, I present different types of credibility practices that were identified across the eight cases. I demonstrate that there are different ways in which model-professionals come to know a model and ascertain its credibility. First, I describe each set individually, after which I will discuss intersections between the sets and explore the power asymmetries that may emerge as a consequence of the time, skills, and resources required to engage in them

Intuitive practices

Not all model-professionals interact with a model directly. The majority of my informants never conducts a model run and learn about the model through intuitive practices. Models may be complicated quantifications, yet many model-professionals ascertain the credibility of a model based on intuition. Other words that my informants used include gut-feeling and common sense. When asking my informants to explain, they described a practice of comparing model mechanics to their understanding of a target system, and model outcomes to those of other sources. If a model corresponds to what they know about a target system, model-professionals will be inclined to trust it. If model outputs are counterintuitive however, that will be of concern. One informant explains:

> Sander: Previously, the number of mortgages was always higher. But for the recent outputs of the WEF model, the conveyance deeds exceeded the number of mortgage deeds. That was quite strange; it was something I had never seen before in my time here. (…) I remember last year, when we used the model, it showed us this change. But we didn't believe it. For every property transaction there had to be a mortgage. So how could the model be right? That year we chose to overrule the model and adjusted the level of conveyance deeds to match the number of mortgages. But in the end, the model was right. Our figures show that the number of mortgages is actually smaller than the number of conveyance deeds.

The WEF model outputs were at odds with Sander's intuition; the model outputs showed something that he considered to be impossible. His colleagues shared this view and concluded the WEF model had to be wrong. They chose to deviate from the model and adjusted the projected workload in accordance with their intuition. However, after taking stock of the actual workload that the Dutch Land Registry had seen that year, they had to conclude that the model had been right. The following year, the WEF model outputs were again on odds with the intuition of the model-professionals. They believed that the residential housing market was starting to recover, while the model predicted ongoing decline. Again, they chose to overrule the model outcomes. This shows the relative importance of intuitive credibility practices. My informants explained that the sources contributing to this intuition range from newspaper articles, to peers of the model-professionals, and outputs of other models. If the outputs of a model are at odds with intuition, this causes conflict; either the model is wrong, or the model-professionals' intuition is amiss. Resolution of this conflict can have two outcomes: the model or its outputs are adjusted; or the model-professional adjusts his intuition to match the model. To resolve the conflict, model-professionals turn to their peers. For instance, if a model-professional finds a particular model output to be counterintuitive he or she may ask his expert colleagues for clarification of that output. One model-professional explains:

> Jack: There are outputs that people are puzzled by, which we try to give a good pounding to find out what has actually happened. When we've got strange results where older cohorts are getting less or more money than the other, it is just testing why things happen and come up. If you got outliers, then you need to know why you've got that, because there are either good reasons for it, or there is something wrong with the modelling.

Amongst model-professionals, the capability to explain model outputs to their peers is important. It is one of the ways in which expert and non-expert model-professionals use to resolve conflicts that may arise from counter-intuitive results.

Formal practices

The intuitive practices outlined above are frugal; comparison of a model output to a newspaper article is fast, as is comparison of the outputs of two models. For the formal practices discussed in the following, this is different. These methods require technical expertise, and more effort on the part of the model professional. One formal credibility practice concerns the comparison of model outputs to observed data. If the model forecasts, for instance, an economic growth of 1% for 2015 in 2014, then in 2015 a model-professional can compare that forecast with the actual economic growth for that year. One informant explains:

> Cor: If the outputs feel off, at a certain point you will question the model. I used historical data. I did the maths for 2005 to 2009. So, I took the predicted number of conveyance deeds and compared that with our actual workload. So even in times of economic growth, the model underestimated the number of deeds. Then, after the shift in economic growth from 2008 to 2009 from the model was too optimistic about the number of deeds. So, in the first instance, the model was too pessimistic and in the second instance it was too optimistic. So, if you do that for the whole period from 2005 to 2009, then you can see that the difference between the model outputs and our actual workload increases

Such validation is useful, yet only allows model-professionals to ascertain the credibility of a model retroactively. Thus, validation is not necessarily useful in the context of policy making, where the credibility of evidence needs to be evaluated more swiftly. Moreover, for models that make predictions decades into the future, such validation can be problematic altogether. Another set of practices can best be described as uncertainty analysis. This refers to a set of methods that allow expert model-professionals to describe the set of possible model outputs and their respective probabilities. One method that is used for uncertainty analysis is sensitivity analysis. Sensitivity analysis determines how susceptible model outputs will be to changes in inputs. This allows a model-professional to get a better feel for the quality of the model outputs. If for instance the model outputs would deviate strongly with a small change to the inputs, this would be cause for caution (Walker et al., 2003). My informants suggested that sensitivity analysis is the preferred method for uncertainty analysis. However, that technique is not applicable to every model type – for some models it may be too time consuming to complete a large number of model runs. In some cases, sensitivity analysis was considered too costly for the purpose of the model.

Procedural practices

Model-professionals also engage in procedural practices. These practices include writing model documentation, checking the model for errors, and keeping track of what the model is. Expert model-professionals engage in these practices to prevent errors from getting into the model undetected and to allow other model-professionals and stakeholders to readily ascertain the credibility of the model. One informant explains:

> Jack: And I think, during this process again the analysts have been clear that we need to do this but we also need to have a period of peer review. I can sort of vaguely remember when the MacPherson report came out and people were thinking what will we do, are we sure that we got the processes in place? Certainly, before I moved across, I got the sense that the lead analyst was putting procedures in place. This is how it looks, obviously we can do something really quickly, we want to make sure that this is sensible, and we have somebody that actually checks the modelling to make sure it has been done properly.

One of the procedural practices that can serve to make a model more credible concerns documenting where the model inputs come from and what changes have been made to the model over the course of its use. Model-professionals write documentation to keep track of changes or make references inside the model itself. Failure to engage in these practices can make model-professionals distrustful of a model. Another set of practices revolves around what can be called the ‘four eyes’ principle, the idea that any change should be reviewed by at least two people. This developer of the Pensim2 model details how the Department for Work and Pensions has put in place procedures and a governance structure that serves to ensure that no errors creep into the model:

> Harry: We do that through the process of formal change requests and problem logs, where there is a single central place where we record what we have done. What the request is which then, one of my team will look at and use it to help inform the change that they are looking to make. And then work with the expert on, to build it into the model. Then yeah, this is a sort of quality assurance process, in which we got a formal template that we fill in. But then there is also this change control process of an actual release, which is subtly different, so all team members are working on different things. We have one big spreadsheet which lists all the modules and all the processes and all the static code. And everyone, for every change, we enter into the sheet, we changed this, this is an entirely new bit, we are going to delete this bit and I use that to try and organize the process in which we built the model. And then, step by step, we will do release 15.04 version 8, all building on top on one of another. And in each stage you get a final sign-off from a user or a developer, more than one, to agree that the modelling is working as intended and its doing what they wanted it to do in the first place and that the results look in some sense sensible.

Procedural practices can be time consuming and some model-professionals view them as overkill and potentially frustrating. Regardless, having an audit trail of versions and input data serves as a signal to non-expert model-professionals that a model is credible.

Evaluative practices

The final set of practices my informants described were evaluative practices. One type of evaluative practices concerns formal reviews. These are in-depth investigations into the credibility of a model, conducted by model-professionals that are external to the group in which the model is used. The formal reviews generally culminate in a report that details the strong and weak points of a model in comparison with other models and offers suggestions on how to improve the model. This developer talks about the organizations that reviewed Pensim2:

> Ben: Part of building that authoritativeness was getting it audited. Way back when. So, we had the Institute for Fiscal Studies look over it and make some recommendations. And we also had the Congressional Budget Office of the United States do an audit of it. So, we haven't done anything as formal as that since then but we feel that puts the model on a secure basis. One of the things I have to consider is whether we should spend time and money to get another review done by somebody of outside just to check that it is still fit for purpose.

Besides such time-consuming formal reviews, model-professionals can also seek feedback external to their model's group by attending conferences or seminars or organizing events themselves. This allows them to interact with other model-professionals who are not involved with the model. These external model-professionals can then scrutinize the model by asking questions and by presenting their own take on how such modelling should be done.

Feedback need not be actively solicited but can also be invited by uploading model documentation on the Internet. Several informants suggested that open access is the pinnacle of transparency and would therefore be an excellent means towards credibility. This allows any model-professional to review how the modelling was done, but also allows for scrutiny by the general public and media. However, it is important to note that experiences with such open-access schemes vary. Of the two cases in which the model is open access, the 2050-calculator and the UKTimes model, model-professionals argued that in itself open access will not be very effective to develop credibility. A developer of the UKTimes model explains:

> Luke: Something that we struggled with -with this little mini project that we did and that report is on - was what transparency means in the context of a model for which there are hundreds of different technologies [*sic*]. Because, it is only really a few people who can understand what is going on. And it being written down, having everything open access is one form of transparency, but you know. Documenting everything in minute detail does not necessarily help everybody understand what is driving the outputs.

So even though open access to a model is perceived as one of the best means to achieve credibility, it may not actually have an impact on credibility. In the case of the UKTimes model, the open access version is not the same as the version that will be used to inform policy. This is due to the sensitive data that is in the policy version of the model. This can lead one to question the usefulness of open access; if the open-access model and the operational version of the model are not the same, this could create a false sense of credibility.

Discussion

Not all model-professionals ascertain, develop, and contest the credibility of a model in the same way; they have different credibility repertoires. Asymmetries in credibility repertoires occur for several reasons: model-professionals may lack the necessary skills to perform a particular practice, the mechanics of a model may not permit particular practices – for instance, the duration of each run may be several hours, thus rendering sensitivity analysis impractical to some, and model-professionals may perceive particular practices unnecessary for performing credibility – for example, if they feel that credibility has already been established beyond doubt. Because of their divergent credibility repertoires, model-professionals enact credibility in different ways. These different enactments need not align; the use of a model output in a certain way can be seen by some model-professionals as credible, but not to others:

> Luke: The MARKAL model is used to generate a key number in the 2007 impact assessment for the climate change act; it was used to generate a cost number for the cost of meeting the climate change act targets. I know at least one person in this building, who is a modeller and who thinks that is an inappropriate use of the tool like that. That it should not be trying to forecast a number like that. So, you have people who are very experienced in these tools thinking that that is an inappropriate policy use, and others thinking it is great.

Amongst expert model-professionals, the conditions under which outputs are considered credible tend to be more explicitly defined. As a result, conflicts between experts can be resolved by reverting to formal credibility practices. However, in situations where experts and non-experts disagree about the credibility of a model, such formal practices will not work towards resolution. The contents of such analysis is unlikely to mean much to someone who is not trained in statistics. However, evidence that such an analysis was conducted can regarded as a sign of credibility. To non-experts, evidence that expert model-professionals have engaged in formal and procedural practices can contribute towards the credibility of a model. In some situations, non-experts can choose to defer judgment to others by means of evaluative practices. Initiation of such evaluation practices may be costly but results in a seal of approval from the organization that conducted the review. Not all asymmetries emerge from differences in how model-professionals enact credibility, some originate from the organizational context. Some credibility practices may not sit comfortably with the demands put on model-professionals by decision makers:

> Ethan: I mean you didn't build the quick and dirty model because that made you happy. You did it because someone needed an answer. The first thing that will happen is that someone says I need an answer and you'll say the problem is I will have to follow these [quality assurance] processes. So, you know I will come back to you in 12 weeks. That's not good enough, I need an answer now.

Through interaction between model-professionals, credibility repertoires may align over time. This does not necessarily mean that all professionals agree on the credibility of a model, but it may entail the emergence of a working balance. Model-professionals seem to invest significant time and resources in enacting model credibility at the start of model development, this effort seems to subside as models age. For example, the Quitsim model replaced a predecessor that overestimated the efficiency of the anti-smoking marketing campaigns conducted by Public Health England. The existing model presented an alternative view on the target system that was at odds with the one presented by the new model, Quitsim. This meant the model-professionals who developed the Quitsim model faced considerable difficulty with getting their model accepted as credible. For older models, model-professionals may have less freedom for adjustments, as one informant explains:

> Lars: There's situations where you have a coalition agreement. Then we quantify that using SAFFIER. So, then you would let SAFFIER work on its own. But at a later time, policies from that agreement might end up in the forecasts. When intended policy gets implemented. Then the outputs have to be consistent to a certain degree.

Models that have been around for quite some time, may have a considerable authority. For example, the SAFFIERII model provides the legal underpinning for the Dutch government's expenditure. When a model holds such an important role in the policy making process, decision makers may be tempted to develop policy that performs well in the model. One informant explains:

> Pieter: The parties try to develop policy alternatives that receive favourable SAFFIERII outcomes. They would all like have the policy alternative with the best reduction in unemployment. In order to do that, they have to understand how their policies are implemented in the model. You might have a brilliant policy idea, but if it doesn't perform well in the model it is no use. In my opinion, if it's a good policy, you should ignore the model altogether. In practice, however, there is a dynamic of parties looking to use the SAFFIERII outcomes to get good publicity. There is no issue with that as long as the model is a good representation of reality.

In the case of SAFFIERII, political parties try to use the model to get a favourable rating for their proposed policies. In a sense, they 'play' the model. This is remarkable, because it suggests that the model may have a profound impact on how parties formulate their policy. In a system where parties need their policies to perform well in SAFFIERII, only those policies that are perceived to work well in the model will get proposed. If SAFFIERII is an accurate model of the Dutch economy, this is of limited consequence. However, as pointed out by other SAFFIERII model-professionals, it may take time to include changes to the Dutch economy into the model.

The credibility practices that model-professionals engage in ensure that the model can be used to inform policy making. When successfully black boxed, models isolate people from complexity and help to make commensurable otherwise complex policy alternatives. This is valuable to those making policy, because it allows them to compare policy alternatives on an equal footing. Yet, this commensurability comes at a price. The black boxed model obscures assumptions, simplifications, and bias from those not familiar with the model mechanics. Those further removed cannot readily contest the credibility of a model and can hardly be sure that the model represents the target system in a way that aligns with their world view. They cannot engage in credibility practices that would let them inspect the model mechanics. As such they cannot see the details of the worldview that is embedded within. Discussions on the model take place between model-professionals and do not involve all those who use – or are affected by – the model's outputs. While model-professionals engage in practices which help them to ascertain the credibility of a model, the lid remains firmly on the black box to those further removed. This means that it is hard for the general public to scrutinize the model.

Concluding remarks

Those working with models devote considerable non-technical effort to make and maintain models as credible sources of evidence to inform policy making. For a new model to inform policy making, it needs to first establish itself as a credible source of evidence. This requires a model to facilitate the divergent credibility repertoires of those that work with. Although ongoing work is necessary to maintain the credibility of a model, the frequency and level of scrutiny to which the black boxes of models are opened subsides as they become more embedded within an organization (see Knorr-Cettina, 1981). The variability of these practices is important because it results in power asymmetries which impact the embedding of a quantification. This presents a challenge because models are continuously updated to reflect changes in the underlying target system. Expert model-professionals can adjust the mechanics of the model in good faith, yet those without the access or means to inspect the model may not notice and are unlikely to comprehend the full impact of such changes if they do. Vice versa, expert model-professionals may not oversee the ways in which a model may come to dictate what is perceived as feasible policy. Four sets of credibility practices were identified:

> **Intuitive practices:** Frugal practices which revolve around comparing the mechanics and outputs of a model to other sources;
>
> **Formal practices:** Practices which involve subjecting the model to one or more analysis techniques to the model;
>
> **Procedural practices:** Practices that encompass the development and maintenance of an audit trail
>
> **Evaluative practices:** Practices which situate a model in relation to others, list its strengths and weaknesses, and provide suggestions.

I demonstrated that model-professionals have different credibility repertoires and in effect enact credibility in different ways. The work done through the credibility practices is necessary for models to fulfil their function in policy making and the credibility of models is as much enacted as the realities that the models help bring about. The divergent credibility repertoires of model-professionals reveal a power asymmetry where arbitrary decisions by expert model-professionals can have a profound impact on the policy making process. I conclude that:

(1) Where many have called for increased transparency (Diakopoulos & Friedler, 2016; Pasquale, 2015), I show that achieving transparency is at best problematic and at worst unattainable for non-experts. The models I studied here were complicated quantifications. The state-of-the-art in data science techniques such as deep learning, however, produces algorithms that are more dynamic and intricate than the models studied here. The expert model-professionals in my case studies went to great lengths to understand model mechanics, yet only a few of them felt they had an exhaustive understanding of the model in question.

(2) The models I studied were black boxes to most model-professionals that worked with them. Despite this lack of transparency, the models were considered credible enough to inform policy. In some, model-professionals suspend some of their disbelief and accept the black boxed model as a shared worldview. In other cases, my informants

suggested that the model was so authoritative that it had a real impact on what types of policy were considered feasible. Arbitrary or accidental decisions made by experts in the development of a model can thus impact society in profound ways. Through this process, models may contribute to bringing about the – stylized- version of a target system that they seek to represent.

(3) Many of the advantages that are attributed to the use of model may not hold in organizational decision-making contexts. The predictive, organizational, and communicative advantages that are associated with the use of quantifications often require direct interaction with the model itself. It is problematic for non-expert model-professionals to become this familiar with a model. Even if they do achieve this level of understanding, the codification of models and the speed at which the model is adapted presents a considerable challenge to keep this knowledge up to date. This makes it practical impossibility for decision makers to augment model outcomes with human judgment in the sense suggested by Goodwin et al. (2007).

My findings are important in relation to recent efforts to regulate algorithms, such as the 'right to explanation' prescribed by the GDPR (Goodman & Flaxman, 2017). Machine learning algorithms in particular are notoriously hard to explain (Mittelstadt et al., 2016), yet this paper demonstrates that even comparably simple models can be hard to grasp to those working with them daily. Although model-professionals often work in close collaboration and over time develop practices to scrutinize a model, they often remain black boxes to all but a few specialists. This suggests that meaningful oversight of machine learning algorithms that transcend our cognitive capacity presents a formidable challenge for those not working with a model on a daily basis. It follows that transparency of algorithms to the general public seems problematic at best given that they may not have the expertise, the time, or inclination to engage with a model. Communication on algorithms to the general public is outside the scope of this paper yet presents an important avenue for future work. In the meantime, those seeking to explain models to the general public would do well to consider the practices through which lay audiences ascertain the credibility of these models may be widely different from their own. This is particularly salient in relation to models that underpin high stakes government decision making.

Notes

1. This paper considers policy making as the process of government to develop interventions which contribute to solving a wide range of perceived societal issues (see Stewart et al., 2007).
2. It is important to stress the model studied here *resemble* hypothetical-deductive machine because not all were developed (see Table 1) adhering strictly to the statistical modelling tradition. Rather, the model developers used a variety of techniques to manufacture the algorithmic models. While the focus was on parametric models and the model developers often worked to ensure that individual variables were significant, non-parametric techniques such as decision trees were also used. It should be clear that the algorithmic models discussed here were not neural networks, a class of algorithmic models that is notoriously complex (Peng et al., 2016).
3. For sake of brevity I refer to algorithmic models as models throughout the paper.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributor

Daan Kolkman is a senior researcher in decision making at the Jheronimus Academy of Data Science. His research revolves around the sociology of quantification and decision support decision systems. Daan received his PhD in sociology from the University of Surrey (England) for his work on computational models in government. He was supervised by Nigel Gilbert, Tina Balke, and Paolo Campo at the Centre for Research in Social Simulation [email: d.kolkman@tue.nl].

ORCID

Daan Kolkman http://orcid.org/0000-0001-9836-0212

References

Barocas, S., & Selbst, A. (2016). Big data's disparate impact. *California Law Review*, *104*(1), 671–729. www.jstor.org/stable/24758720

Bissell, C., Dillon, C., & Mansnerus, E. (2012). The inner world of models and its epistemic diversity: Infectious disease and climate modelling. In G. Gramelsberger & E. Mansnerus (Eds.), *Ways of thinking, ways of seeing. Automation, collaboration, & e-services* (pp. 167–195). Springer. https://doi.org/10.1007/978-3-642-25209-9

Boulanger, P. M. P.-M., & Bréchet, T. (2005). Models for policy in sustainable development: The state of the art and perspectives for research. *Ecological Economics*, *55*(3), 337–350. https://doi.org/10.1016/j.ecolecon.2005.07.033

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Cardon, D., Cointet, J. P., & Mazieres, A. (2018). Neurons spike back: The invention of inductive machines and the artificial intelligence controversy.

Cash, D., & Clark, W. C. (2001). From science to policy: Assessing the assessment process. *John F. Kennedy School of Government Faculty Research Working Papers Series*.

Christin, A. (2017). Algorithms in practice: Comparing web journalism and criminal justice. *Big Data & Society*, *4*(2), 1–14. https://doi.org/10.1177/2053951717718855

Diakopoulos, N., & Friedler, S. (2016). How to hold algorithms accountable. *MIT Technology Review*, *17*(11).

Diez, E., & McIntosh, B. S. (2011). Organisational drivers for, constraints on and impacts of decision and information support tool use in desertification policy and management. *Environmental Modelling & Software*, *26*(3), 317–327. https://doi.org/10.1016/j.envsoft.2010.04.003

Dourish, P. (2016). Algorithms and their others: Algorithmic culture in context. *Big Data & Society*, *3*(2), 1–15. https://doi.org/10.1177/2053951716665128

Espeland, W. N., & Stevens, M. L. (2008). A sociology of quantification. *European Journal of Sociology*, *49*(3), 401–436. https://doi.org/10.1017/S0003975609000150

Floridi, L., & Taddeo, M. (2016). What is data ethics? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2083), 20160360. https://doi.org/10.1098/rsta.2016.0360

Gilbert, N., & Mulkay, M. (1984). *Opening Pandorra's box*. Cambridge University Press.

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, *38*(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

Goodwin, P., Fildes, R., Lawrence, M., & Nikolopoulos, K. (2007). The process of using a forecasting support system. *International Journal of Forecasting*, *23*(3), 391–404. https://doi.org/10.1016/j.ijforecast.2007.05.016

Head, B. W. (2010). Reconsidering evidence-based policy: Key issues and challenges. *Policy and Society*, *29*(2), 77–94. https://doi.org/10.1016/j.polsoc.2010.03.001

Heuts, F., & Mol, A. (2013). What is a good tomato? A case of valuing in practice. *Valuation Studies*, *1*(2), 125–146. https://doi.org/10.3384/vs.2001-5992.1312125

Kemper, J., & Kolkman, D. (2019). Transparent to whom? No algorithmic accountability without a critical audience. *Information, Communication & Society*, *22*(14), 2081–2096. https://doi.org/10.1080/1369118X.2018.1477967

Knorr-Cetina. (1981). *The manufacture of knowledge: Towards a constructivist and contextual view of science*. Pergamon.

Kolkman, D. (2020). The usefulness of algorithmic models in policy making. *Government Information Quaterly*. https://doi.org/10.1016/j.giq.2020.101488

Kolkman, D. A., Campo, P., Balke-Visser, T., & Gilbert, N. (2016). How to build models for government: Criteria driving model acceptance in policymaking. *Policy Sciences*, *49*(4), 489–504. https://doi.org/10.1007/s11077-016-9250-4

Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago press.

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. Harvard University Press.

Law, J., & Mol, A. (2006). The actor-enacted: Cumbrian Sheep in 2001. In C. Knappett & L. Malafouris (Eds.), *Material agency* (pp. 57–77). Springer US.

Mazzotti, M. (2017). Algorithmic life. Retrieved February 12, 2020, from https://lareviewofbooks.org/article/algorithmic-life/

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 68–82. https://doi.org/10.1177/2053951716679679

Pasquale, F. (2015). *The black box society*. Harvard University Press.

Patton, M. Q. (2002). Purposeful sampling. In *Qualitative evaluation and research methods* (Vol. 3). Sage.

Peng, G., Ritchey, N. A., Casey, K. S., Kearns, E. J., Privette, J. L., Saunders, D., & Ansari, S. (2016). Scientific stewardship in the open data and big data era – roles and responsibilities of stewards and other major product stakeholders. *D-Lib Magazine*, *22*(5-6), 1–14.

Pielke, R. A. (1999). Who decides? Forecasts and responsibilities in the 1997 red river flood. *Applied Behavioral Science Review*, *7*(2), 83–101. https://doi.org/10.1016/S1068-8595(00)80012-4

Shapin, S. (1995). Cordelia's love: Credibility and the social studies of science. *Perspectives on Science*, *3*(3), 255–275. http://nrs.harvard.edu/urn-3:HUL.InstRepos:3293019

Stewart Jr., J., Hedge, D. M., & Lester, J. P. (2007). *Public policy: An evolutionary approach*. Nelson Education.

Strauss, A. L., & Corbin, J. M. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Routledge.

Taylor, L. (2017). What is data justice? The case for connecting digital rights and freedoms globally. *Big Data & Society*, *4*(2), 205395171773633. https://doi.org/10.1177/2053951717736335

Treasury, H. M. (2013). Review of quality assurance of government analytical models: Final report.

van Daalen, C. E., Dresen, L., & Janssen, M. A. (2002). The roles of computer models in the environmental policy life cycle. *Environmental Science & Policy*, *238*, 1–11. https://doi.org/10.1016/S1462-9011

Van der Sluijs, J. P. (2002). A way out of the credibility crisis of models used in integrated environmental assessment. *Futures*, *34*(2), 133–146. https://doi.org/10.1016/S0016-3287(01)00051-9

van Doorn, N. (2017). Platform labor: On the gendered and racialized exploitation of low-income service work in the 'on-demand' economy. *Information, Communication & Society*, *20*(6), 898–914. https://doi.org/10.1080/1369118X.2017.1294194

Venkatesh, V., & Bala, H. (2008). Technology acceptance model 3 and a research agenda on interventions. *Decision Sciences*, *39*(2), 273–315. https://doi.org/10.1111/j.1540-5915.2008.00192.x

Venturini, T., Jensen, P., & Latour, B. (2015). Fill in the gap. A new alliance for social and natural sciences. *Journal of Artificial Societies and Social Simulation*, *18*(1), 1–4. https://doi.org/10.18564/jasss.2729

Walker, W. E., Harremoës, P., Rotmans, J., van der Sluijs, J. P., van Asselt, M. B. A., Janssen, P., & Krayer von Krauss, M. P. (2003). Defining uncertainty: A conceptual basis for uncertainty management in model-based decision support. *Integrated Assessment*, *4*(1), 5–17. https://doi.org/10.1076/iaij.4.1.5.16466