The role of trust in the use of artificial intelligence for chemical risk assessment

Wassenaar, P.N.H.; Minnema, J; Vriend, J.; Peijnenburg, W.J.G.M.; Pennings, J.L.A.; Kienhuis, A.

Regulatory Toxicology and Pharmacology 148 (2024) 105589

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

journal homepage: www.elsevier.com/locate/yrtph

The role of trust in the use of artificial intelligence for chemical risk assessment

Pim N.H. Wassenaar [a,*], Jordi Minnema [a], Jelle Vriend [a], Willie J.G.M. Peijnenburg [a,b], Jeroen L.A. Pennings [a], Anne Kienhuis [a]

[a] National Institute for Public Health and the Environment (RIVM), P.O. Box 1, 3720 BA, Bilthoven, the Netherlands
[b] Institute of Environmental Sciences (CML), Leiden University, P. O. Box 9518, 2300 RA, Leiden, the Netherlands

ARTICLE INFO

Handling Editor: Dr. Lesa Aylward

Keywords:
In silico toxicolog
Artificial intelligence
Machine learning
Chemical risk assessment
Trust
Regulatory acceptance

ABSTRACT

Risk assessment of chemicals is a time-consuming process and needs to be optimized to ensure all chemicals are timely evaluated and regulated. This transition could be stimulated by valuable applications of *in silico* Artificial Intelligence (AI)/Machine Learning (ML) models. However, implementation of AI/ML models in risk assessment is lagging behind. Most AI/ML models are considered 'black boxes' that lack mechanistical explainability, causing risk assessors to have insufficient trust in their predictions.

Here, we explore 'trust' as an essential factor towards regulatory acceptance of AI/ML models. We provide an overview of the elements of trust, including technical and beyond-technical aspects, and highlight elements that are considered most important to build trust by risk assessors. The results provide recommendations for risk assessors and computational modelers for future development of AI/ML models, including: 1) Keep models simple and interpretable; 2) Offer transparency in the data and data curation; 3) Clearly define and communicate the scope/intended purpose; 4) Define adoption criteria; 5) Make models accessible and user-friendly; 6) Demonstrate the added value in practical settings; and 7) Engage in interdisciplinary settings. These recommendations should ideally be acknowledged in future developments to stimulate trust and acceptance of AI/ML models for regulatory purposes.

1. Introduction

Risk assessment of chemicals is necessary to ensure safe production and use. However, this process is time-consuming, and under pressure due to the large number of chemicals being placed on the market (Wang et al., 2020) combined with the limited capacity of risk assessors (European Commission, 2019). Accordingly, there is a need to optimize and improve the risk assessment process. A number of factors are important to accurately perform chemical risk assessment. These factors include availability of reliable methods, and availability of a substantial amount of data for the specific chemicals (European Commission, 2020; ECHA & European Commission, 2019).

The use of Artificial Intelligence (AI) – and Machine Learning (ML) – can stimulate the transition towards a more efficient risk assessment, as it allows more efficient use of the available data and knowledge. The development and use of AI/ML methods has evolved rapidly in the last decade, particularly due to an increase in available data (Pawar et al., 2019) and computational capacity (OpenAI, 2018). These methods can specifically be used to develop *in silico* AI/ML models, which could potentially be used supplementary or as an alternative to current risk assessment practices.

Although there are numerous developments in the field of *in silico* toxicology, the direct application and implementation of AI/ML models in risk assessment are lagging behind (Lin et al., 2022; Rivetti et al., 2023). Several factors can be identified that hamper the acceptance and use of AI-based models in risk assessment. The vast amount of data generated in toxicology, often referred to as 'big data', might pose a risk of getting overwhelmed by the amount of data, losing sight of the objective of the hazard or risk assessment to be undertaken (Richarz, 2019). Developments in big data sources have also advanced computational toxicology, including AI/ML methods. As a result, more and more AI/ML algorithms and approaches are becoming available, making it increasingly difficult to achieve consensus on what model would best fit a specific hazard or risk assessment purpose (Lin et al., 2022).

* Corresponding author.
E-mail address: pim.wassenaar@rivm.nl (P.N.H. Wassenaar).

https://doi.org/10.1016/j.yrtph.2024.105589
Received 17 August 2023; Received in revised form 26 January 2024; Accepted 21 February 2024
Available online 23 February 2024

Additionally, most automated models based on AI are considered 'black boxes' that lack mechanistical explainability, causing users (i.e., risk assessors) to have insufficient trust in their predictions. Technical challenges for the application of AI/ML models in hazard and risk assessment are related to AI/ML model performance, integration of data inputs and interpretation of data outputs. An overall challenge is gaining trust, not only on the technical aspects, but also on beyond-technical aspects of AI/ML models. Considering trust as a combination of technical and beyond-technical aspects will ultimately define the confidence in the reliability of AI/ML models that can be used for risk assessment.

The aim of this study is to explore 'trust' as an essential factor to gain regulatory acceptance of AI/ML models. As recommended in a recent JRC report, "Chemicals regulation needs robust trust practices, that allow for critical interrogation and shared understandings of data, evidence and methods" (JRC, 2022). Here we attempt to objectify this 'trust' for use of AI/ML models in risk assessment by opening the 'black box': 'How can trust be defined?', and 'At what levels does trust play a role for the use of AI/ML models in the chemical risk assessment?'. By increasing our understanding of what is important from a risk assessment perspective, we can improve the applicability of newly developed AI/ML models and stimulate the discussion between computational modelers and risk assessors on this topic: when risk assessors can express their concerns and specify the essential criteria, computational modelers can focus on these elements to increase the applicability of *in silico* models for regulatory purposes. This step is essential to stimulate future optimization and improvements of the chemical risk assessment with AI/ML models.

We first provide some background on the use of *in silico* models in risk assessment (section 2.1) and on AI/ML methods (section 2.2). Next, we more elaborately describe the elements of trust (section 3.1), and specifically highlight the elements that are essential to risk assessors (section 3.2). As an input for and as part of this research we organized a workshop with chemical risk assessors and computational modelers from the Dutch National Institute for Public Health and the Environment (RIVM, n = 27), covering risk assessors with expertise in human toxicology and ecotoxicology, and a wide variety of regulatory frameworks (including REACH, Water Framework Directive, and Pesticides, Biocides and Pharmaceuticals regulations). The results of this workshop are incorporated as a main source of information for section 3.2 and the program of the workshop is available as Supplementary Material Table S1. The paper closes with several recommendations for risk assessors and computational modelers to stimulate the transition and acceptance of the use of AI-based predictive models for regulatory purposes (section 4).

2. Background

2.1. In silico models in risk assessment

The development and application of *in silico* models for chemical risk assessment is not new, and various models have already been implemented and applied in several regulatory frameworks for decades. This is nicely presented in a timeline of the application of *in silico* AI/ML models, recently published by Lin and Chou (2022), Hemmerich and Ecker (2020) and several others, describing the application of *in silico* models already as early as 1890 (Lin et al., 2022; Hemmerich et al., 2020). Most of the currently applied models in chemical risk assessment focus on hazard identification using classification on potency (e.g., predicting categories of low, moderate and high toxicity) or regression-based models (e.g., predicting LC50 in mg/L). In addition, most models focus on general toxicity/apical endpoints, though some focus on mechanisms of action. To a somewhat lesser extent *in silico* exposure models are used, which are mainly based on mechanism-based systems. Generally, available models are mainly applied (and applicable) for screening and prioritization for hazard identification rather than being directly applied to characterize risks (e.g., for the derivation of Environmental Quality Standards or Human Reference Doses).

Table 1
Illustration of various types of *in silico* models (following Hemmerich and Ecker, 2020), including some examples that are frequently applied for screening and risk assessment.

Types of *in silico* methods	Examples
Expert-based systems: models that are created using the knowledge and experience of experts to deduce or explain toxicity (mechanisms)	Structural alerts Read-across
Mechanism-based systems: models that are created using mechanism-based processes that are derived from scientific principles and assumptions	Exposure and multimedia fate models PBK (Physiologically based kinetic)/ TKTD (toxicokinetic-toxicodynamic) models Adverse outcome pathways (AOP)-based models
Learning-based systems: models that are created by learning patterns from data and parameters for the model	Machine learning models, including quantitative structure activity relationships (QSAR)

Different types of *in silico* methods can be distinguished, including expert-based, mechanism-based and learning-based systems (Lin et al., 2022; Hemmerich et al., 2020). In Table 1 examples of these model types are provided, in which it should be noted that there can be many exceptions to this subdivision. For example, structural alerts could also be defined and generated with learning-based systems (Valsecchi et al., 2019), and various elements of a read-across could be performed by learning-based systems (e.g., automated chemical similarity searches (Gadaleta et al., 2020; Vigano et al., 2022; Wassenaar et al., 2022)). Furthermore, it could be discussed whether exposure models are part of the field of *in silico* toxicology, which mainly focuses on computational models to predict hazards. Nevertheless, we have included these sub-groups in this overview for completeness.

Particularly the last category, the learning-based systems, is well known, as it includes the so-called Quantitative Structure Activity Relationships (QSAR) models. It is this category that also includes the increasingly popular machine learning models.

2.2. Artificial intelligence methods in risk assessment

Artificial Intelligence is not new, but has been developing rapidly due to increasing computing capacity, an increase in the amount of data, and an increase in insights into how to apply these methods for societal challenges. Artificial Intelligence broadly refers to any human-like behavior (or intelligence processes) displayed or simulated by a machine or computer system (Pérez Santín et al., 2021). Machine learning is a subset of AI which allows a machine to automatically learn patterns and relations from data (Pérez Santín et al., 2021; Janiesch et al., 2021). Accordingly, the well-known QSAR models can be considered as an example of machine learning, as they also learn patterns from past observations (Hemmerich et al., 2020).

ML models are increasingly developed within the field of computational toxicology. For the development of a ML model commonly three components are needed: *i)* features or descriptors (i.e., input data), *ii)* target data (i.e., examples of desired output data), and *iii)* an algorithm which relates the input features to the desired output. Various types of ML models are being developed and most focus on predicting potential hazardous effects of chemicals using supervised ML models, and thereby try to fill in parts of the hazard and risk assessment. In addition, there are also developments that focus more on a facilitating function of AI to contribute to the risk assessment process rather than on performing (steps of) a risk assessment (Bersani et al., 2022; OpenAI, 2023; van de Schoot et al., 2021). More details about AI and its use in predictive toxicology and chemical risk assessment are described elsewhere (Lin et al., 2022; Hemmerich et al., 2020; Wang et al., 2021; Wittwehr et al., 2020), as well as descriptions of useful data sources, tools and platforms (Pawar et al., 2019; Gozalbes et al., 2018).

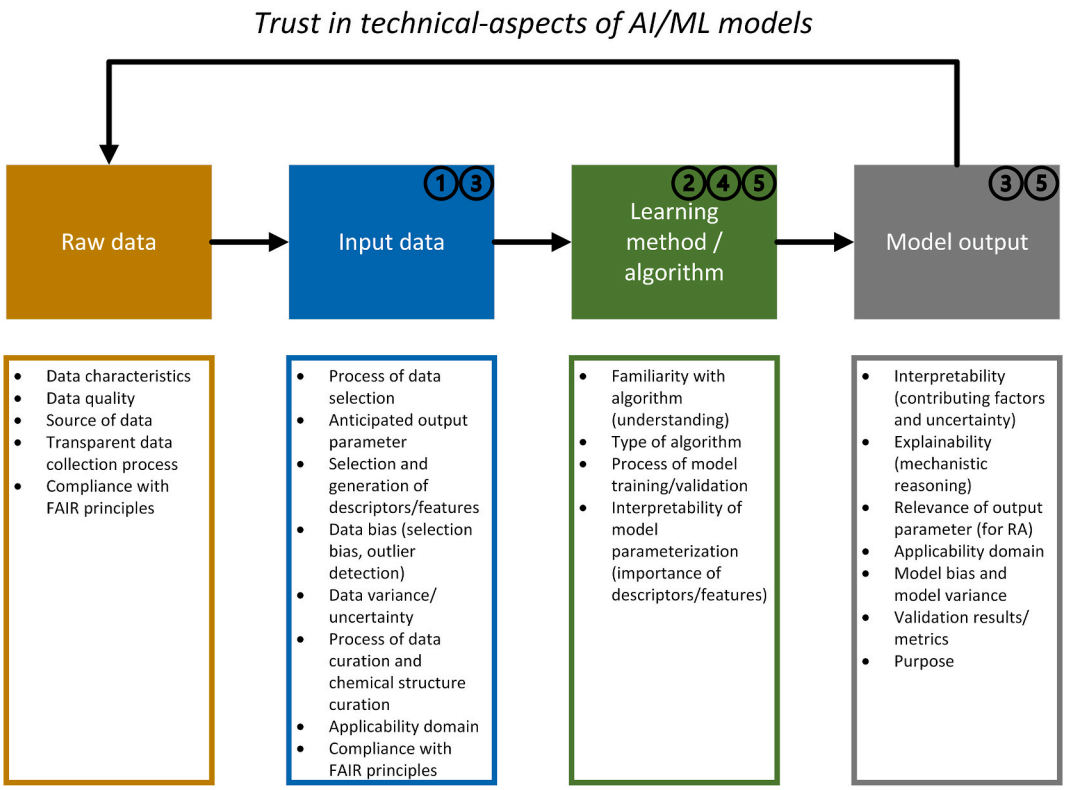

Fig. 1. Overview of different levels at which trust in technical aspects of AI/ML models could be affected, visualized in a typical machine learning pipeline: from raw data to input data, learning method/algorithm and model output. For these levels, we list several elements that could potentially influence the level of trust in AI/ML models (adopted from Chatzimparmpas et al., 2020). The feedback loop indicates that constant feedback is required to identify and optimize those elements that are important to build trust in technical aspects of AI/ML models. The encircled numbers represent the relation with the OECD QSARs principles (OECD, 2004), and at which level they are mainly involved (1 = a defined endpoint; 2 = an unambiguous algorithm; 3 = a defined domain of applicability; 4 = appropriate measures of goodness-of-fit, robustness, and predictivity; 5 = mechanistic interpretation, if possible). Abbreviations: FAIR = Findable, Accessible, Interoperable and Reusable; RA = risk assessment.

The advantage of ML models over other types of models lies in their ability to *learn* which features in the input data are relevant to predict a desired outcome. Simple ML models with low computational cost – like classical QSAR models – may solely focus on linear trends between the input features and the target data, whereas more complex models may also incorporate non-linear and/or abstract correlations that are difficult to be extracted by humans (Wang et al., 2021). Because data is needed to learn, the application of (traditional) ML models is particularly valuable when dealing with large and complex datasets. Furthermore, as the complexity of a ML model increases (e.g., deep learning and multi-task predictions (Hemmerich et al., 2020)), the data requirements to fully leverage the added value of such a model also increase.

On the other hand, a general downside of ML models is the common difficulty of interpretability and transparency of model predictions. In addition, there is a risk of 'overfitting', which refers to a model that predicts the training dataset very accurately, but is not able to generalize sufficiently to other data (Lin et al., 2022). The true predictive performance of a model for the broader chemical universe is difficult to determine, and links/contributes to the interpretability of model predictions. Quite often, ML models are described as 'black-boxes' (or 'closed boxes') (Lin et al., 2022). This statement may not always be fully substantiated, and the definition of this terminology may vary between persons and can be based on diverse underlying arguments. The elements related to the interpretability and transparency of model predictions are important for imbedding such models for various applications in the chemical risk assessment, and will be further discussed in this paper.

3. Role of trust

3.1. Trust in artificial intelligence based models

Trust is a key factor for acceptance of any method in risk assessment by stakeholders (JRC, 2022). There are various definitions of trust, which are also closely related to definitions of scientific credibility: "the willingness of persons to base decisions on information obtained from the model" (Schruben, 1980). This section is based on earlier definitions of trust in relation to artificial intelligence (Chatzimparmpas et al., 2020; Siau and Wang, 2018), and is slightly adapted to the context of our research/field. Accordingly, we assume that "the level of trust a person has in someone or something can determine that person's behaviour" and "trust can also define the way people interact with technology" (Siau and Wang, 2018).

Gaining trust is not a static process, but is rather dynamic and it requires time to build trust. Building trust starts with initial trust formation followed by continuous trust development. Continuous trust will depend on the performance and purpose of AI technology. This also shows its vulnerability; continuous trust development can easily ('in seconds') be hampered once AI technology fails to perform and does not fit the purpose (Siau and Wang, 2018). Although these aspects of trust were described for AI technology in general, they are also applicable to the use of AI/ML models for application in risk assessment.

Various types of trust can be defined that are all involved throughout these two different phases (i.e., build initial trust and continuous trust development), and are all influencing one's level of trust in AI/ML models. Here these types of trust are roughly divided into 'trust in the technical aspects of AI/ML models' and 'trust in beyond-technical

aspects of AI/ML models', although there can be some overlap between these aspects. Both types of trust are present – and influence each other – and contribute to the overall perception of trust in AI/ML models.

3.1.1. Trust in technical aspects of AI/ML models

Trust in the technical aspects of AI/ML models includes the elements where risk assessors or computational modelers mainly think of when talking about trust in AI/ML models, such as quality of input data, applicability domain, or model performance. Trust in technical aspects entails elements that contribute to the understanding of how a specific AI-based model works.

Several 'starting criteria' have already been set by the OECD in 2004 for the use of QSARs for regulatory purposes (OECD, 2004), and the criteria remain applicable to current models. These criteria can serve as a starting point to which AI/ML models should conform to in order to build trust in the technical aspects of AI for risk assessment. The OECD guidelines demand regulatory acceptable models that conform with the following principles:

1. a defined endpoint,
2. an unambiguous algorithm,
3. a defined domain of applicability,
4. appropriate measures of goodness-of-fit, robustness, and predictivity,
5. mechanistic interpretation, if possible.

These principles have recently been extended with assessment elements to verify whether a prediction is reliable in the context of a specific application (OECD, 2023).

When extending these technical aspects from the OECD guidelines to the AI/ML model development pipeline, four levels of technical trust can be identified (see Fig. 1), including trust in: raw data, input data, AI/ML learning method (i.e., the algorithms), and the model output. These levels are interconnected, and several elements are listed in Fig. 1 that could potentially affect the trust in AI/ML models at various levels of the machine learning pipeline. Lack of trust in any of these aspects will add up and together contribute to the total lack-of-trust perception in an AI/ML model. Furthermore, these aspects not only contribute to initial trust formation, but are also important to stimulate continuous trust development by constantly improving AI/ML models. Accordingly, it is important to get a better understanding of the elements that matter most to risk assessors, so that these elements can specifically be considered during model development (see Section 3.2).

It is recognized that the perception of the importance of the different elements listed in Fig. 1 to build trust in technical aspects of an AI/ML model will likely vary between persons, e.g., based on experience (as further discussed under 'Trust in beyond-technical aspects'). However, similarities are to be expected within stakeholder-groups with regard to their perception of what are essential technical specifications. For instance, this may differ between risk assessors (i.e., domain experts) and machine learning experts. Furthermore, the perception of the importance of the various elements may depend on the intended risk assessment purpose for which the model is used. For instance, AI/ML models intended for screening and prioritization are often used to identify chemicals for further evaluation and not to enforce specific regulatory or risk management measures. As such, different or less elements (see Fig. 1) might be essential to build trust in technical aspects of AI/ML models for screening and prioritization compared to elements considered important for models directly used for regulatory decision making. In addition, the importance of the elements may differ between regulatory frameworks within which a model is applied.

3.1.2. Trust in beyond-technical aspects of AI/ML models

Besides the technical aspects influencing trust, which are expected to play at a stakeholder-group level, societal and personal related elements also influence trust in AI/ML models. These elements are referred to as trust in beyond-technical aspects, and include personal/societal elements such as personality and culture. Beyond-technical aspects are influenced by both a person's general beliefs as well as by model-specific elements (as are the technical related aspects).

General beliefs are influenced by someone's personality, and are not related to a specific AI/ML model. For instance, the personality or disposition to trust could be thought of as the general willingness to trust others, which in turn depends amongst others on different experiences and personality types. Apart from personality, someone's general beliefs

Table 2
Overview of beyond-technical aspects that contribute to trust in AI/ML models for risk assessment, both during initial trust formation and continuous trust development. The aspects are adapted from Siau and Wang (2018). Note that some elements have a strong connection with trust in technical aspects of AI/ML models.

Initial trust formation		
Level	Aspects	Description (examples)
Performance	Representation	- Elaborate and clear description and documentation of an AI/ML model
	Image/perception	- Perception, understanding and trust of AI in general - Perception of the trustworthiness of the source that developed the AI/ML model
	Reviews from other users	- Availability of critical evaluation by peers - Quality of the critical evaluation by peers (e.g., the amount of negative/positive reviews for AI/ML models by peers)
Process	Transparency	- Availability of open source publication of the AI/ML model's input data, output data and algorithms
	Trialability	- Opportunity for end-users to try the AI/ML model (e.g., more resistance is to be expected for a new technology from people who cannot use it)
Continuous trust development		
Level	Aspects	Description (examples)
Performance	Usability and reliability	- Intuitiveness and availability of the AI/ML model (an AI/ML model that is difficult to use, with a high downtime will not be popular for users)
	Collaboration and communication	- Possibility for continuous interaction with developers and end-users (e.g., computational modelers, risk assessors, commercial parties, etc.) (in case of AI/ML models for risk assessment this is required to evaluate whether the method still adds to the purpose of reliable risk assessment)
	Security and privacy protection	- Guarantee of data integrity - Protection of privacy and confidentiality of related input and/or output data of AI/ML models
	Training and education	- Teaching and explaining how AI/ML models operate, how they can be used for a specific purpose and how the data should be interpreted
Purpose	Level of application	- Understanding and appreciation of the specific purpose for which the AI/ML model is developed and applied (is the AI/ML model developed for screening or for full risk assessment?; is the AI/ML model developed to substantiate/prove an effect or is it developed to prove the absence of an effect?)
	Goal congruence/ consistency	- Meeting expectations, aligning with purposes (making sure that AI's goals are congruent with human goals is a precursor in maintaining continuous trust) - Ensuring ethical and legal frameworks to be developed for use of AI/ML models in risk assessment

are also influenced by a person's direct surroundings such as institutional factors and cultural background (Siau and Wang, 2018). Institutional factors include conditions and formal and informal rules in society that constrain behavior. Cultural background can be based on ethnicity, race, religion, and socioeconomic status, and can be associated with a country or a particular region. The impact of someone's general beliefs on trust in any new technology will be roughly similar regardless of the technology. For instance, a person with a high-trusting stance would be generally more likely to accept AI/ML models, as well as other new technologies.

Besides the here described 'general beliefs', there are also model-specific 'beyond-technical' attributes that contribute to someone's trust in new developments. This level of trust will be different across models and can be considered as model-related characteristics. Although some of these model-related characteristics have a commonality with technical-related aspects that influence trust in AI/ML models, the aspects described here are more influenced by behavioral and social elements and/or considerations rather than the more quantifiable and objectifiable elements that are part of the technical-related aspects that influence trust. The model-specific 'beyond-technical' attributes can be analyzed from three perspectives, including i) the performance, ii) its process and iii) its purpose; and can be separated into various aspects that apply to the different phases of building trust (i.e., initial trust formation and continuous trust development). An overview of potentially relevant model-specific 'beyond-technical' attributes that may contribute to someone's trust in AI/ML models for risk assessment is presented in Table 2. The overview follows the division by Siau and Wang (2018), applied to trust in AI/ML models for risk assessment.

3.2. Relevance for risk assessment

To gain more insight into the elements in Fig. 1 and Table 2 that are considered most important by risk assessors, we organized a workshop with computational modelers and risk assessors (details of this workshop are in the Supplementary Material Table S1). During the workshop, there was a general agreement among the risk assessors that AI/ML methods are an emerging and promising technology in chemical risk assessment. Accordingly, there was a general feeling that we need to investigate the options for implementation of AI/ML models in risk assessment practices. From this workshop, several conclusions were drawn that are important for trust and adoption of AI/ML models by risk assessors, which are described below. Some of these conclusions focus more on technical aspects and criteria for AI/ML models (e.g., point 1 and 2), whereas others focus more on elements that influence the general regulatory acceptance of AI/ML models (e.g., points 3–6), covering both technical and beyond-technical aspects.

1. Insights into underlying data:

The participants of the workshop identified the underlying data and its properties (chemical domain, distribution/ranges of parameters, test conditions, etc.), as one of the most important elements when considering whether to use an AI/ML model. They noted that transparency of the training data is essential, and that training data should be subjected to strict quality criteria to avoid training the model on unreliable data. In addition, the applied criteria for evaluating the data should be transparent and clearly communicated. Furthermore, selection bias in the training data should be considered, as the available data is not always fully representative of the task at hand.

2. Interpretability and understandability:

The interpretability of models was raised as an important element when considering the acceptance of AI/ML models for use in risk assessment. However, it was interesting to observe that the risk assessors used different definitions of 'interpretability' as the discussion progressed during the workshop. For example, several participants mentioned that a risk assessor should have insight into the model's applicability domain as well as the model's measure for uncertainty in the predicted outcome to allow for appropriate interpretability. Other participants mentioned that risk assessors should be able to understand and interpret the model in a similar way as they are able to understand and interpret current tools used for risk assessment, e.g., through detailed documentation in guidance documents. However, at the same time it was noted that it is not yet clear what criteria should be met for a satisfactorily interpretable AI/ML model for risk assessment. To improve interpretability, a model should at least have a relation with the biological or ecological process or mechanism it predicts. In addition, to stimulate interpretability, a clear description of limitations of a model was considered very important, leading to the question: 'For what purpose/situation is the AI/ML model not appropriate and should it not be used?'. As different purposes and frameworks for risk assessment have different criteria and/or requirements, computational modelers should cater to these needs in collaboration with risk assessors in order to develop models that can be trusted and accepted by risk assessors.

3. Appreciation by regulators and other stakeholders:

Appreciation of AI/ML models by regulators and other stakeholders is an essential element for trust and adoption in risk assessment. To appreciate an AI/ML model, risk assessors need proof that the model performance (relevance and reliability) has been assessed in relation to the intended purpose. In order to ensure that model performance fits with the intended purpose for risk assessment, collaboration between computational modelers and risk assessors is needed already at the level of model development, as also mentioned in '2'. Appreciation and consensus among computational modelers and risk assessors on suitability of AI/ML outcomes to be used as part of decision-making in risk assessment is key for adoption. True implementation would also include further steps such as acceptance of AI-generated outcomes by other stakeholders (e.g., risk managers, industry), regardless of whether the result is a negative or positive outcome.

4. Availability of existing data and methods:

Another important conclusion was that the availability of existing data or methods for specific risk assessment purposes heavily influences the acceptance of data predicted by AI/ML models. If no existing data or methods are available, risk assessors tend to trust/accept predicted data from AI/ML models more easily. If existing methods or models are available (and already applied in practice), risk assessors tend to compare the performance of a new AI/ML models to this current application (considering whether it outperforms the existing method for the task at hand). However, the possibility that the existing method might be less-suited is often not being considered, hampering the opportunity of the AI/ML model to show added value in risk assessment decision making.

5. Trialability of AI/ML models:

A good way to increase trust in a model is to let risk assessors work with an AI/ML model themselves, instead of solely relying on the expert computational modelers. An easy-to-use AI/ML model would enable risk assessors to learn about the tool and to compare its results to a conventional risk assessment procedure, in which the point raised under '4' should be kept in mind. This would require development of useable and understandable AI/ML models that are available to all end-users. If the tool works adequately and to their satisfaction, this will ultimately increase the appreciation and trust of the risk assessor in the AI/ML model and the potentially ultimate acceptance of the tool. In addition, it was mentioned that a proof-of-concept (i.e., a clear application of the model for a real-life example(s) that illustrates its usability and limitations)

Table 3
Recommendations to users (i.e., risk assessors) and developers (i.e., computational modelers) of AI/ML models to build trust and stimulate regulatory acceptance for future developments.* = see section 3.2. ** = risk assessors, computational modelers, or both.

#	Recommendation	Related conclusion (s)*	Explanation	Main responsibility**
1	Keep models simple and interpretable	2	Computational modelers tend to implement the absolute best statistical performing model, which can be difficult to interpret and can lead to a 'black-box' feeling among risk assessors. Often, simpler models offer comparable statistical performances, while being much more explicable (i.e., higher 'regulatory performance'). Hence, models should be as simple as possible and as complex as needed, in which more time should be invested into explaining and clarifying what the model does. To further increase interpretability the applicability domain should be clearly defined, preferably quantitatively, in order to ensure proper use of the AI/ML models. Furthermore, guidance should be provided to risk assessors, regulators and/or other stakeholders to correctly interpret the model predictions in relation to the applicability domain.	Computational modelers
2	Offer transparency in the data and data curation	1	The data used to develop AI/ML models is key to the model performance. In order to increase trust into a model's performance, the underlying data as well as the curation steps should be FAIR, transparent and well-explained.	Computational modelers
3	Clearly define and communicate the scope/intended purpose	2	The intended purpose of the model (e.g., screening vs risk assessment) should be clearly communicated, as the model is also developed and validated for that application. In addition, it should also be clarified in which situations the use of the model is not appropriate.	Computational modelers
4	Define adoption criteria	4	Regulatory adoption criteria should be (loosely) defined. By critically evaluating the requirements from a risk assessment perspective, the essential elements will become more clear. Distinctions could be made between the intended purpose of the models and possibly the various regulatory frameworks. In addition, this may also depend on whether there is already a standard non-AI/ML-based approach implemented. When defining criteria, we should be critical/careful when comparing AI/ML models to current standards, and we should be aware of the added value and dangers of such comparisons.	Risk assessors
5	Make models accessible and user-friendly	5	Risk assessors should not depend on computational modelers to run the model. By making models accessible and user-friendly, risk assessors can use the models themselves and develop a feeling for the model's behavior and predictions, which can in turn be beneficial in increasing trust.	Computational modelers
6	Demonstrate the added value in practical settings (i.e., practical validation)	3, 4, 6	AI/ML models are often studied in research settings, where the outcomes are known and the performance can be quantified. However, in order to adopt AI/ML	Both

(*continued on next page*)

Table 3 (*continued*)

#	Recommendation	Related conclusion (s)*	Explanation	Main responsibility**
			models in general risk assessment, their added value needs to be demonstrated outside of research settings. For example, one can perform a case study/proof-of-principle where AI-based risk assessment is performed besides conventional risk assessment. Depending on the case, AI/ML models may speed up the assessment, or offer new insights. By showing the value of AI/ML models in the everyday work of risk assessors, trust can be gained.	
7	Engage in interdisciplinary settings	3, 6	For the development and use of AI/ML models in risk assessment various disciplines are involved, including data and AI-experts (i.e., computational modelers) and domain-experts (i.e., risk assessors). To increase the level of trust in this kind of models we need to understand each other and learn to speak each other's language. Accordingly it is important to invest time in interdisciplinary activities to build a common understanding of what is possible and what is required.	Both

would also help to build trust in the use of an AI/ML model. This can for example be done by performing case studies together with stakeholders.

6. Adding human intelligence to AI/ML models:

Some participants mentioned that the final decision should always be taken by human assessors, and should not be taken over by the computer. This may also help with solving legal issues, as there is not (yet) a status of AI as natural person or legal entity in the existing legal framework. Therefore, the ultimate risk assessment decision must remain in the hands of (human) risk assessors.

4. Conclusions and recommendations

Predictive AI/ML models are being developed at a rapid pace and represent promising methods to innovate and advance current chemical risk assessment practices. One way to stimulate this transition is to facilitate regulatory acceptance of AI/ML models that contribute to reliable risk assessment. This study focused on elements of trust that are important for risk assessors in order to adopt AI/ML models for acceptance in regulatory frameworks, and was investigated through a workshop.

Based on this study we propose the recommendations as highlighted in Table 3, which include the following elements of trust that are important for risk assessors in order to trust an AI/ML model: transparency, interpretability, understanding, appreciation, documentation, trialability and usability. These recommendations, which relate to earlier validation principles (OECD, 2004; OECD, 2023) and credibility factors (Patterson et al., 2021), can serve as a starting point for both, risk assessors and computational modelers, and should ideally be acknowledged in future developments to stimulate trust and regulatory acceptance of AI/ML models.

Funding and declaration of interests

This work was supported by funding from the RIVM data utilization program.

CRediT authorship contribution statement

Pim N.H. Wassenaar: Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing. **Jordi Minnema:** Conceptualization, Methodology, Writing – review & editing. **Jelle Vriend:** Conceptualization, Methodology, Writing – review & editing. **Willie J.G.M. Peijnenburg:** Methodology, Writing – review & editing. **Jeroen L.A. Pennings:** Methodology, Writing – review & editing. **Anne Kienhuis:** Conceptualization, Methodology, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

We are extremely grateful to the workshop facilitator, Caroline Moermond, and all of the participants of the workshop for their valuable contributions.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.yrtph.2024.105589.

References

Bersani, C., et al., 2022. Roadmap for actions on artificial intelligence for evidence management in risk assessment. EFSA Support. Publ. 19 (5).
Chatzimparmpas, A., et al., 2020. The state of the art in enhancing trust in machine learning models with the use of visualizations. Comput. Graph. Forum 39 (3), 713–756.
ECHA & European Commission, 2019. REACH Evaluation Joint Action Plan. Ensuring Compliance of REACH Registrations.
European Commission, 2019. Findings of the Fitness Check of the Most Relevant Chemicals Legislation (Excluding REACH) and Identified Challenges, Gaps and Weaknesses.

European Commission, 2020. Chemicals Strategy for Sustainability: towards a Toxic-free Environment.
Gadaleta, D., et al., 2020. Automated integration of structural, biological and metabolic similarities to improve read-across. ALTEX 37 (3), 469–481.
Gozalbes, R., Vicente de Julián-Ortiz, J., 2018. Applications of chemoinformatics in predictive toxicology for regulatory purposes, especially in the context of the EU REACH legislation. International Journal of Quantitative Structure-Property Relationships 3 (1), 1–24.
Hemmerich, J., Ecker, G.F., 2020. *In* silico toxicology: from structure-activity relationships towards deep learning and adverse outcome pathways. Wiley Interdiscip. Rev. Comput. Mol. Sci. 10 (4), e1475.
Janiesch, C., Zschech, P., Heinrich, K., 2021. Machine learning and deep learning. Electron. Mark. 31 (3), 685–695.
JRC, 2022. Addressing Evidence Needs in Chemicals Policy and Regulation.
Lin, Z., Chou, W.C., 2022. Machine learning and artificial intelligence in toxicological sciences. Toxicol. Sci. 189 (1), 7–19.
OECD, 2004. OECD Principles for the Validation, for Regulatory Purposes, of (Quantitative) Structure-Activity Relationship Models.
OECD, 2023. (Q)SAR Assessment Framework: Guidance for the Regulatory Assessment of (Quantitative) Structure – Activity Relationship Models, Predictions, and Results Based on Multiple Predictions.
OpenAI, 2018. AI and compute. https://openai.com/research/ai-and-compute.
OpenAI, 2023. ChatGPT. https://openai.com/blog/chatgpt.
Patterson, E.A., Whelan, M.P., Worth, A.P., 2021. The role of validation in establishing the scientific credibility of predictive toxicology approaches intended for regulatory application. Comput Toxicol 17, 100144.
Pawar, G., et al., 2019. *In* silico toxicology data resources to support read-across and (Q) SAR. Front. Pharmacol. 10, 561.
Pérez Santín, E., et al., 2021. Toxicity prediction based on artificial intelligence: a multidisciplinary overview. WIREs Computational Molecular Science 11 (5).
Richarz, A.-N., 2019. Big data in predictive toxicology: challenges, opportunities and perspectives. In: Issues in Toxicology.
Rivetti, C., Campos, B., 2023. Establishing a NexGen, mechanism-based environmental risk assessment paradigm shift: are we ready yet? Integrated Environ. Assess. Manag. 19 (3), 571–573.
Schruben, L.W., 1980. Establishing the credibility of simulations. Simulation 34 (3), 101–105.
Siau, K., Wang, W., 2018. Building trust in artificial intelligence, machine learning, and robotics. Cut. Bus. Technol. J. 31 (2), 47–53.
Valsecchi, C., Grisoni, F., Consonni, V., Ballabio, D., 2019. Structural alerts for the identification of bioaccumulative compounds. Integrated Environ. Assess. Manag. 15 (1), 19–28.
van de Schoot, R., et al., 2021. An open source machine learning framework for efficient and transparent systematic reviews. Nat. Mach. Intell. 3 (2), 125–133.
Vigano, E.L., et al., 2022. Virtual extensive read-across: a new open-access software for chemical read-across and its application to the carcinogenicity assessment of botanicals. Molecules 27 (19).
Wang, Z., Walker, G.W., Muir, D.C.G., Nagatani-Yoshida, K., 2020. Toward a global understanding of chemical pollution: a first comprehensive analysis of national and regional chemical inventories. Environ. Sci. Technol. 54 (5), 2575–2584.
Wang, M.W.H., Goodman, J.M., Allen, T.E.H., 2021. Machine learning in predictive toxicology: recent applications and future directions for classification models. Chem. Res. Toxicol. 34 (2), 217–239.
Wassenaar, P.N.H., Rorije, E., Vijver, M.G., Peijnenburg, W., 2022. ZZS similarity tool: the online tool for similarity screening to identify chemicals of potential concern. J. Comput. Chem. 43 (15), 1042–1052.
Wittwehr, C., et al., 2020. Artificial Intelligence for chemical risk assessment. Comput Toxicol 13, 100114.