AI & SOCIETY (2023) 38:5–20
https://doi.org/10.1007/s00146-021-01282-1

ORIGINAL ARTICLE

Principle-based recommendations for big data and machine learning in food safety: the P-SAFETY model

Salvatore Sapienza[1] · **Anton Vedder**[2]

Received: 22 January 2021 / Accepted: 2 September 2021 / Published online: 11 October 2021

Abstract

Big data and Machine learning Techniques are reshaping the way in which food safety risk assessment is conducted. The ongoing 'datafication' of food safety risk assessment activities and the progressive deployment of probabilistic models in their practices requires a discussion on the advantages and disadvantages of these advances. In particular, the low level of trust in EU food safety risk assessment framework highlighted in 2019 by an EU-funded survey could be exacerbated by novel methods of analysis. The variety of processed data raises unique questions regarding the interplay of multiple regulatory systems alongside food safety legislation. Provisions aiming to preserve the confidentiality of data and protect personal information are juxtaposed to norms prescribing the public disclosure of scientific information. This research is intended to provide guidance for data governance and data ownership issues that unfold from the ongoing transformation of the technical and legal domains of food safety risk assessment. Following the reconstruction of technological advances in data collection and analysis and the description of recent amendments to food safety legislation, emerging concerns are discussed in light of the individual, collective and social implications of the deployment of cutting-edge Big Data collection and analysis techniques. Then, a set of principle-based recommendations is proposed by adapting high-level principles enshrined in institutional documents about Artificial Intelligence to the realm of food safety risk assessment. The proposed set of recommendations adopts Safety, Accountability, Fairness, Explainability, Transparency as core principles (SAFETY), whereas Privacy and data protection are used as a meta-principle.

Keywords Big data · Machine learning · Food safety · Risk assessment · Data governance · Data ownership

1 Introduction, research question and methodology

Ensuring access by all people to safe, nutritious and sufficient food falls within the remit of UN Sustainable Development Goal (SDG) number 2. In the European Union (EU), where most people do not experience hunger, food safety represents a major concern due to its crucial role in promoting human health and fostering the internal market.[1] The European Food Safety Authority (EFSA) is the EU institution responsible for food safety risk assessment. In the last few years, EFSA has shown a growing interest towards Machine Learning Techniques (MLTs). The deployment of MLTs is made possible by multiple factors: the abundance of high-quality data available for scientific analysis, the presence of data warehouses for gathering and

✉ Salvatore Sapienza
salvatore.sapienza@unibo.it

Anton Vedder
anton.vedder@kuleuven.be

[1] University of Bologna, Bologna, Italy

[2] Faculty of Law, Centre for IT & IP Law, KU Leuven, Leuven, Belgium

[1] Recital 1 of the General Food Law Regulations (Regulation (EC) No 178/2002, GFLR) states that"[t]he free movement of safe and wholesome food is an essential aspect of the internal market and contributes significantly to the health and well-being of citizens, and to their social and economic interests.

structuring information, data standards, and sufficient computational power to generate predictions, simulations, and other inferences.

The use of MLTs entails a significant paradigm shift in food safety risk assessment, with serious implications for public health in the EU. Risk assessment covers areas such as regulated products (e.g. plant protection products, GMOs), food packaging, food contact materials, and other regulated areas (Van der Meulen 2013), hence its social relevance for citizens and undertakings. The unprecedented ability to classify and predict risks through Big Data implies that probabilistic methods can be used alongside traditional deterministic methodologies. While trials and experimental tests are still relevant, technological advances are shaping a future in which the importance of food safety data analysis is becoming higher than the observation of real and tangible samples. This trend is consistent with other transitional domains that, taken together, constitute the so-called 'mangrove society' (Floridi 2018). In the course of its de-materialisation, however, datafied and data-centric food safety generates both opportunities and risks.

On the one hand, the presence of artificial agents that cooperate with the scientist in predicting future risks is likely to foster the "great promise" of UN SDGs and EU policy goals that everyone in Europe will benefit from a timely and accurate risk assessment, while the industry will be satisfied by faster and 'smart' procedures. On the other hand, MLTs can generate an additional layer of distrust in risk assessors to the one highlighted by the Fitness Check on the General Food Law Regulation.[2] In 2021, a significant amendment known as the Transparency Regulation was made to the EU food safety legislative framework to promote openness and transparency in risk assessment. The rationale underlying this piece of legislation is to promote trust in risk assessors by expanding EFSA obligations to disclose the information on which its scientific evaluation is grounded.

Despite these legislative efforts, trust may remain a significant issue. Literature on the ethics of artificial intelligence (AI) has suggested that trust in the algorithm is a crucial component of the social acceptability of MLTs (Taddeo and Floridi 2018). In the novel Regulation, no provisions clearly regulate the use of Big Data and MLTs. The technical protection of confidential data coming from industry, the transparency of analytical methods and the processing of personal data for risk assessment purposes have been partly overlooked or delegated to other acts, thus generating confusion on the applicable legal regime. As a result, data governance and data ownership represent two major sources of concern and uncertainty. Regrettably, complementary solutions are often hindered by the conflicting interests of heterogeneous stakeholders, including the food industry, data providers, academic communities independently studying food and feed hazards, and the general public (Sapienza and Palmirani 2018).

To briefly restate the motive of this study, food safety risk assessment activities fall within the scope of the EU law and this legislation has been recently amended to increase trust in risk assessors. However, the peculiarities of emerging analysis techniques raise questions about the resilience of this Regulation to this novel trends. If not addressed properly, the misuse of MTLs by the competent authorities might vanish the efforts to increase trust and prevent the social acceptability of these techniques.

When confronting with unregulated technological trends, ethics can provide guidance to the actors subject to incomplete norms (Floridi 2018). The relationship between ethics and any existing normative frameworks can be expressed—following Floridi's interpretation of the role of ethics in this debate—either as a challenge to the existing legal framework to be used in a *de iure condendo* perspective ("hard ethics") or as what ought and ought not to be done over and above the existing regulations ("soft ethics"). In the latter sense, scholars (Morley et al. 2020a) have identified an agreement on ethical principles reached in the communities that have published documents on ethics of AI (e.g., the AI4People initiative, the Asilomar conference,[3] IEEE[4]). However, these frameworks are yet to be applied and adopted in real-world applications. Peculiarities of food safety risk assessment include heterogeneous data sources and information types, the presence of several stakeholders, the variety of MLTs whose adoption is foreseen. Taken together, they make up an ideal set of conditions to adapt high-level ethical guidelines to the real world. In this study, we investigate what principles should steer data ownership and data governance solutions in the current and foreseeable future of food safety risk assessment.

Therefore, this paper aims to conceptualise an introductory ethical framework to integrate, interpret and align the technical and legal domains of food safety risk assessment to the principles guiding MLTs and AI deployment that have emerged in the academic literature and policy documents. On their bases, this research proposes as a set of principles to support decision-making processes related to data and algorithms in the domain at stake, ultimately contributing to the trustworthiness of risk assessment practices.

[2] EU Commission, 'Refit Evaluation of the General Food law (Regulation (EC) No 178/2002' SWD (2018) 38 final. The initiative has been launched as a reply to the citizens' legislative initiative "Ban Glyphosate and Protect People and the Environment from Toxic Pesticides" that followed the Monsanto Paper scandal. Taken together, these facts highlight the lack of trust on risk assessment activities and the scepticism towards the validity and independence of their results.

[3] Future of Life Institute, "Principles developed in conjunction with the 2017 Asilomar conference".

[4] *IEEE Standards v2* "Ethically aligned design".

To fulfil its goals, this short paper adopts a cross-disciplinary methodology grounded on three pillars. First, principles enshrined in the model should be grounded on the existing legal framework, consistently with the soft ethics approach. Moreover, the state of the art in MLTs applied to food safety risk assessment has to be kept into account. These conditions are also necessary to avoid misconceptions in what is technically feasible or legally required or admitted (*ad impossibilia nemo tenetur*). Second, to prevent arbitrariness and partiality (Vedder 2019a), a selection of ethical documents and, within them, candidate principles should be listed according to pre-defined criteria. AI ethics charters and documents released by the industry and institutions have recently proliferated (Hagendorff 2020) and, in principle, all would deserve equal attention. However, given the centrality of EFSA (an EU institution) in the domain at stake, this paper only considers ethical papers coming from EU institutions or Member States governmental bodies or appointed entities identified as such by reviewers (Fjeld et al. 2020; Jobin et al. 2019; Morley et al. 2020a). Then, principles have been selected in light of their proximity to the domain at stake, i.e., data analysis practices implemented within the risk assessment framework and the EU food safety legal framework. To ease their understanding, the discussion on principles has also been enriched by references to the existing literature in the field of AI ethics. Third, the drafting of explicit technical recommendations or legislative proposals is out of the scope of this research due to the lack of proven methods to translate principles into practice (Mittelstadt 2019). Instead, the model and its recommendations should be considered as a set of broad normative statements to support decision-making.

This paper is divided into seven sections. Following this introduction, Section 2 concerns technical advances in food safety risk assessment, whereas Section 3 discusses how data are regulated under the EU regulatory system. Section 4 qualitatively evaluates findings emerging from the technical and legal analysis. Section 5 elaborates the SAFETY—Security, Accountability, Fairness, Explainability, Transparency—model by aligning high-level principles to our domain following a qualitative assessment of the literature about each of them. Section 6 discusses the inclusion of Privacy as a meta-principle for the model according to the same methodology. Finally, a conclusion summarises the study and presents, discusses our findings, and sets directions for further research.

2 Big data, MLTs and the future of risk assessment

Risk assessment is a segment of risk analysis that primarily concerns the uncertainty surrounding the effects of food and feedstuff on human health. The goal of EFSA is to provide scientific opinions to the risk managers (the EU Commission and the Parliament) responsible for the administrative measures related to authorisations, food banning, monitoring programmes and so forth. Scientific opinions are freely available at EFSA Journal.

Food safety risk assessment consists of four steps (Gilsenan 2015):

1. Hazard identification aims to identify negative health effects (e.g., carcinogenicity) that may be caused by the exposure to a particular agent. This step mainly consists in the review of the scientific literature on hazards, such as the active substance of a chemical product.
2. Hazard characterisation measures the relationship between a certain level of exposure and the occurrence of negative health effects, e.g., carcinogenicity.
3. Risk characterisation measures the concrete level of exposure by identifying the level of hazard in food eaten in a given area/time/population.
4. Exposure assessment relies on hazard characterisation and exposure assessment data to predict how likely a certain risk scenario will materialise.

In practice, risk assessment procedures aim to discover how likely risks to human health can occur as a consequence of behaviours related to food farming, production, ingestion, etc. To perform this analysis, it is necessary to process several information-types: results from scientific trials, chemical information related to certain products (e.g., pesticides), individual food consumption data and other information needed to identify and classify dietary habits. Since EFSA has no laboratories, most of such data originate from third parties, including EU Member States (which, in turn, gather data either from the industry or from their own research), applicants in the field of regulated products supporting their claims, independent studies and spontaneous data providers. Data warehousing (EFSA 2015a) and standardisation are consolidated trends (EFSA 2015b).

MLTs are made possible thanks to this infrastructure, and they are likely to affect all the four steps mentioned above. Several technical reports showing the potential of MLTs in food safety risk assessment have been released by EFSA (Cavalli et al. 2019; IZSTO et al. 2017; Jaspers et al. 2018; Naydenova et al. 2019). These comprehensive reviews show that MTLs can be useful throughout all the aforementioned steps of risk assessment. First, scientific literature on a given topic can be automatically reviewed to extract the most relevant papers for the identification of hazards (Step 1). Second, classification tools can identify the extent to which agents can cause negative health effects (Step 2). Third, clustering can be useful to segment countries to identify more exposed consumers in risk characterisation (Step

3). Finally, predictive models can support exposure assessment by generating predictions related to outbreaks (Step 4).

Moreover, the industry can, in turn, use MLTs to generate evidence to be submitted to the Authority. This entails that EFSA itself has to confront both well-known issues in data sharing, including the technical protection of industry-generated data and information regarding individuals, and new concerns specifically related to MLTs, such as the inexplicability of the logic underlying certain algorithms or the trustworthiness of the statistical model and its results.

3 Data transparency and confidentiality in the EU food safety legal framework

Statistical models result from heterogeneous data. Article 33 of the GFLR mandates EFSA to "search for, collect, collate, analyse and summarise relevant scientific and technical data in the fields within its mission", in particular, with regard to food consumption and the exposure of individuals to risks related to the consumption of food, incidence and prevalence of biological risk, contaminants in food and feed, and residues. The composite nature of such information is reflected into the applicable legal regimes. This section aims to clarify their scopes and their interplay by differentiating the personal/non-personal nature of the data at stake.

Food consumption data and background information fall within the scope of the General Data Protection Regulation (GDPR)[5] and Regulation 2018/1725[6] for being linked to identifiable individuals.[7] Food consumption data are usually gathered in a 2-day non-consecutive 24-h food dietary recall, i.e., an individual survey intended to gather data about the food and beverages consumed in the previous 24 h (EFSA 2014). Background information, including age, sex, body weight and height, is collected to identify dietary patterns. Moreover, pre-defined dietary habits, whether through personal choice (e.g., vegetarians) or because of health conditions (e.g., diabetes or coeliac disease) are also recorded if made explicit by the surveyed individual.

As regards non-personal information collected in compliance with Article 33 GFLR, Article 38(1)(c) sets a general obligation "to make public without delay the information on which its opinions are based". However, the publication of this information is limited by the confidential treatment that data providers can request (Article 39(1)). In general, it can be observed that the regulatory framework pertaining to information at stake has been focusing on striking a balance between the need of granting access to data by EFSA, including via Member States, and protecting commercial interests of third parties (namely, the industry) that might be harmed if data were disclosed to competitors. The legal doctrine discussing data ownership related to data made available to EFSA has stated in multiple occasions that the Authority should not qualify as data owner despite the transmission, thus not preventing the applicability of legal instruments such as intellectual property rights, confidentiality rules or contractual limitations to protect commercially sensitive data (Kocharov 2009; Lodge 2003; Simpson 2016). However, the 2019 reform[8] entered into force in March 2021 broadens the scope of information covered by mandatory publication. The list now includes "scientific data, studies and other information supporting applications, including supplementary information supplied by applicants" (new Article 38(d)) and "the information on which [EFSA] scientific outputs, including scientific opinions are based". In both cases, the Authority has to take into account the protection of confidential information that can be requested by originator and safeguards for personal data.

The reform further clarifies the boundaries of the immediate publication of the information mentioned above. New Article 38(1)(a) states that the publication should be without prejudice to (a) "any existing rules concerning intellectual property rights which set out limitations on certain uses of the disclosed documents or their content and (b) any provisions set out in Union law protecting the investment made by innovators in gathering the information and data supporting relevant applications for authorisations ("data exclusivity rules"). Further clarifications state that the disclosure to the public of this information does not grant any license for third parties' use, reproduction or exploitation of the published data (New Article 38(2)).

The scope of confidentiality requests is now restricted to selected items for which confidential treatment can be requested (new Article 39). For data providers, it is now necessary to attach a "verifiable justification that demonstrates how making public the information concerned significantly harms" commercial and marketing interests. In any case,

[5] Regulation (EU) 2016/679. It worth mentioning that issues pertaining to dietary intake information were discussed by Article 29 Working Party's Letter to Paul Timmers' (Annex on health data in apps and devices) when the question regarding the sensitive/non-sensitive qualification of such data was raised. It follows that, if linked to an identified or identifiable individual as in the case at stake, food consumption data can qualify as personal data at least.

[6] Regulation (EU) 2018/1725. This Regulation applies when EFSA qualifies as data controller for determining the purposes of data processing.

[7] Metadata used by EFSA to harmonise food consumption data (available at https://zenodo.org/record/1215993#.XxagrC2Q1TY, accessed on 11/07/2021) show the attribution of unique subject identifiers to surveyed individuals, together with 27 other information-types that might allow the reidentification of the individual. Other scholars (Alemanno and Gabbi 2016, p. 32) also discuss food consumption information in terms of personal data.

[8] Regulation (EU) 2019/1381 OJ L 231/1.

EFSA has to make public the non-confidential version of the submitted dossier without delay.

Furthermore, the category of mixed data can be relevant when personal and non-personal data are within the same dataset and correlate each other.[9] An assessment on the interdependency of personal and non-personal data within the same dataset would also be needed, as this might trigger the provisions enshrined in Article 2(2) of the Free Flow of Non-Personal Data Regulation.[10] If the personal and non-personal data in the dataset are 'inextricably linked' data protection rights and obligations apply to the whole dataset. In the domain at stake, this assessment depends on the context (for instance, food consumption data might be fully anonymised before the insertion in the dataset, which, in turn, does not become mixed). However, the Free Flow of Non-Personal Data Regulation only applies to contracts signed with data processing service providers, as Article 2(3) only restricts its scope to the 'outsourced' operations of public bodies. Vice versa, EFSA 'in-house' data processing activities are not covered by the Free Flow of Non-Personal Data Regulation.

Findings from this section highlight how the existing Regulation sets obligations to manage only few of the several issues related to the novel technical trends in food safety. While setting clear rules to manage the confidentiality of data submitted by the industry and the protection of personal data, the existing framework does not provide adequate guidance for the use of MLTs in this context.

4 Mapping the issues of novel trends in food safety risk assessment

The previous section has described the essential elements of the EU food safety legal framework that are relevant to personal and non-personal data processing. The legislative goal of the new reform was to increase the trust in the risk assessors by introducing rules that mandate the disclosure of data on which risk evaluations are based. However, algorithms are places outside the legislative discourse despite being one the best ways to 'make sense' of data, in particular those large and heterogenous. This section aims to identify concerns emerging from the adoption of MLTs in relation to information-types processed for risk assessment purposes. As noted by scholars in the field of data ethics (Floridi and Taddeo 2016), the discussion of data-related emerging concerns shall be carried out by adopting a perspective that encompasses data, algorithms and practices.

Food consumption and demographic data can be used to generate inferences regarding (a) the health status, thanks to the combination of raw food consumption data, weight and height, and other indicators (Lazarou et al. 2012); (b) religious believes or ethnic origin, as many religions prescribe the observation of fasting (e.g., daylight hours during the month of Ramadan in the Islamic calendar), dietary temporary limitations (e.g., Friday, Ash Wednesday and Good Friday in Catholic Canon Law) or strictly defined rules (e.g., Kashrut dietary laws exclude animals listed in the 613 commandments); (c) philosophical believes, such as veganism and vegetarianism; (d) political opinions, according to certain studies showing some degree of correlation between drinking habits and certain political ideologies (Yakovlev et al. 2013) or describing left-wing (environmentalist, organic and 'farm-to-fork') positions *vis-á-vis* conservative (frozen food, massive portions and energy drink) eating habits (Kazutoshi 2018).

These information-types qualify as 'special categories of data' for the purposes of the data protection law (Art. 9 GDPR, Art. 10 Regulation 2018/1725). However, these sensitive attributes can only be inferred rather than processed as they were provided by the data subject unless they are mentioned in the dietary surveys described in Section II. Dietary intake data can serve thus as a proxy variable for sensitive attributes inferred from food consumption patterns[11], thus raising data protection concerns as regards the secondary processing of these data by the industry and safeguard measures to be implemented to protect them.

It is worth noticing that Member States adopted jeopardised data protection measures when building national food consumption datasets to be included in EFSA's comprehensive database: Spain simply treated personal data "as confidential" (Marcos et al. 2016, para 2.6); France had to obtain a mandatory authorisation by the French Data Protection Authority (Dubuisson et al. 2017, para 2.6); the Netherlands did not report about data protection aspects (Dutch National Institute Public Health 2018, para 2.6); Italy stated that "[t]he survey was exclusively observational and non-invasive, ethical aspects were related only to the collection of information on food habits that may be related to health and thus might be sensitive. INRAN is part of the National Statistical System (SISTAN) and guarantees individual data

[9] The 2019 EU Commission Guidance on the Regulation on a framework for the free flow of non-personal data in the European Union states that "a mixed dataset consists of both personal and non-personal data. Mixed datasets […] are common because of technological developments".

[10] Regulation (EU) 2018/1807 OJ L 303/59.

[11] It worth mentioning that the Article 29 Working Party's Letter to Paul Timmers' (Annex on health data in apps and devices) adopts food consumption data as an example of proxy variable between 'regular' and 'special' categories of data. The example is further discussed elsewhere (Malgieri and Comandè 2017b).

protection. An additional ethical committee review of the study protocol was considered unnecessary" (Sette et al. 2011, p. 923). Diverse data protection measures might be symptomatic of difficulties in applying data protection laws to food consumption information. Such difficulties might arise from the data aggregation that usually takes place when risk assessment activities are performed.

Jeopardised and sub-optimal compliance shall also be read in the light of EFSA's 'probabilistic turn'. The shift from deterministic methods of analysis to the probabilistic modelling made possible by the use of MLTs entails the generation of results that only show significant patterns rather than causation. Some authors have referred to these results as "inconclusive evidence" (Mittelstadt et al. 2016). In one of EFSA's case studies (EFSA 2019) on the evaluation of the effects of pesticide on the thyroid over time, the comparison of deterministic and probabilistic models produced nearly identical forecasts of exposure to possible hazards. The observed differences were attributed to the random effect of probabilistic modelling (EFSA 2019, p. 32) that is correlated to the stochastic nature of MLTs. Although these encouraging findings show the potential of cutting-edge data analysis techniques in key areas of risk assessment, new kinds of uncertainties arise due to the probabilistic nature of MLT-generated models. Deterministic algorithms are characterised by an equality relationship such that, given the same input and a functioning software, the output will always be the same. Instead, probabilistic modelling is characterised by results expressed in terms of likelihood. Due to their own nature, MLTs always come with the possibility of an avoidable statistical error (e.g., false positives/false negatives).

It might be the case that probabilistic models are more appropriate to describe real-world phenomena thanks to their adaptability and flexibility. Deterministic algorithms are less adaptive to new trends and always need pre-defined instructions, hence the opportunity of improving the efficiency of risk assessment activities thanks to MLTs.

However, significant concerns related to the probabilistic nature of these algorithms highlighted by the literature deserve attention. First, the correctness of deterministic results can be evaluated by analysing if input data have been validated and the algorithm has correctly executed all the planned instructions. In contrast to deterministic scenarios, probabilistic errors can occur despite the use of high-quality input data and correct functioning of the software. While the new EU reform has endorsed the principle of transparency in risk assessment, there is no specific obligation to make available to the public the method of analysis or the algorithms used either by EFSA when drafting its scientific opinions or by the industry when submitting data.

This is somewhat concerning also in the light of the 'black box' issue (Pasquale 2015). It consists of the "inscrutability" of the logic followed by the algorithms when delivering their outputs (Mittelstadt et al. 2016) and the related possibility of identifying the entity accountable for the decisions taken on their basis. The discussion towards the algorithmic black box has been raised in the context of automated decision-making, the GDPR and, in particular, art. 22 and the contested existence of a 'right to explanation' in the Regulation (Wachter et al. 2017, Malgieri and Comandè 2017a, 2017b). Instead, given the inapplicability of art. 22, the question of how uncertainty should be conceptualised when opaque, inexplicable or unexplainable MLTs are deployed in the public sphere is open.

Furthermore, algorithmic biases represent another source of potential concern. The impact of algorithmic biases on individuals is well documented by scholars discussing the social implications of 'big data' trends (Barocas and Selbst 2016; Mittelstadt et al. 2016; O'neil 2016). These studies have to framed into our domain to evaluate the impact of intake information processing. As we noted, food consumption data are gathered from different individuals that might be classified according to their dietary habits. When considering aggregated food patterns (i.e. the aggregation of food preferences of multiple individuals), at least two kinds of biases can emerge: data-driven bias (e.g., overestimation or underestimation of certain groups/group characteristics) and similarity bias (e.g., when using previous literature as a resource of risk assessment in a manner that reinforce the bias displayed in the previous research).

In addition to the limited applicability of data protection law, biases related to the use of MLTs can originate from the joint processing of personal and non-personal data. Their mash-up raises concerns due to the cumbersome interaction of different legal regimes governing their ownership. In particular, the stratified legal layers of protection cover proprietary data and include confidentiality measures from food law, copyright and *sui generis* rules from the Database Directive,[12] as well as other legal instruments such as trade secrets or contractual age. They might be an obstacle for independent reviewers willing to check the correctness of input data and algorithms, as with in the *Hautala* case.[13]

To briefly restate the main findings emerging from this section, a set of challenges linked to the implementation of MLTs by risk assessors has been identified. Despite being aligned with the research problems that the cited scholars are addressing in related fields, the deployment of MLTs in our research scenario has peculiarities that are strictly linked to the domain at stake. On the one hand, possible

[12] Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases [1996] OJ L 77/20.

[13] Case T-329/17 Heidi Hautala and Others v European Food Safety Authority [2019] ECLI:EU:T:2019:142.

inaccuracies—including inscrutable evidence and biases—are particularly significant because they are grounded on information that reveal sensitive traits of individuals' personality. Such sensitivity is not only justified by the "specific protection" granted by the GDPR to these information-types,[14] but also by their proximity to personal identity, that is—according to certain theories (Tamò-Larrieux 2018; Lynskey 2014; Floridi 2013)—the object of protection of data protection law itself. However, the GDPR and Regulation 2018/1725 are regrettably inapplicable to the data processing at stake: data aggregation techniques, in practice, might prevent *tout court* the applicability of these pieces of legislation.[15] Even if applicable, article 22 of the GDPR and article 24 of Regulation 2018/1725 would not mitigate the concerns related to the use of MLTs over food consumption data as no automated decision-making processes take place in our domain. On the other hand, the nature of probabilistic risk assessment raises questions per se on the accuracy of the results and the possibility of lawfully cross validating the outputs of MLTs. While desirable, such cross validation is unlikely to occur in practice due to the legal means used to protect data and algorithms.

Notwithstanding the limited direct impact over individuals, risk assessment activities are still relevant for their social and collective implications for human health. These are particularly significant also for certain groups that share ethical, philosophical and religious believes also expressed in the form of dietary habits. Accountability frameworks for the negative outcomes of the deployment of MLTs shall also be discussed to identify the entity responsible for incorrect decisions. The next section discusses a theoretical governance model aiming at mitigating these issues.

5 The "SAFETY" Model

The previous sections have identified gaps in the current food safety legal framework as regards the use of algorithms and, in particular MLTs. The lack of explicit provisions on how to use them in our context might originate risks for individuals due to the collective and social consequences of food safety risk assessment for human health. Given the importance of protecting personal and non-personal data against unlawful exploitation and the need of validating scientific results emerging from available data and MLTs, it is necessary to set principles to govern data analysis in a way that it guarantees the respect of ownership rules while preventing biases and immunities in the risk assessors for the use of probabilistic models. Consistently with our research methodology and the 'soft ethics' approach, such principles shall be meant as a complement to the existing regulatory framework.

The SAFETY—Security, Accountability, Fairness, Explainability, Transparency—model discussed below proposes a non-exhaustive and non-hierarchical list of principles that integrate, interpret and align the existing regulatory framework on data ownership and governance to foster a trustworthy use of data and MLTs in EU food safety risk assessment. These elements might be key drivers in the adoption of MTLs in this domain, as theoretical (Cowls et al. 2019) and empirical research (Shin 2021a) suggests, also with regards to other technologies e.g., blockchain (Hwang 2020). This list of principles has been drafted by relying on well-established principles at an EU-level (European Commission 2018) and Member States[16] as interpreted by prominent authors.

Notable examples of principle-based ethical frameworks include the FATE – fairness, transparency, accountability, explainability—model (Shin 2021a), which corresponds to a set of *desiderata* discussed in the current literature of AI ethics (Ferrario et al. 2019). Notably, the correlation between the FATE model and trust AI development and deployment has been suggested in theory and tested empirically, e.g., in the contexts of algorithmic media user recommender systems (Shin 2020, 2021b). Scholars conducting surveys as regards the acceptability of AI decisions noted that "[b]ased on the FATE model, we can infer that trust is closely embedded with these issues as it plays a key role in developing user confidence and credibility. Only when users are assured that their data and privacy are secured with FATE, the user trust is unfolded, and users are willing to provide more of their data to AI" (Shin 2021a).

As anticipated in the methodological section, principles validated by the literature and enshrined in the aforementioned institutional charters have to be adapted to the domain at stake, i.e. agri-food safety assessment. To do so, principles that have reached a certain degree of consensus have been identified. Then, those that should be deemed incompatible with the domain at stake for technical or legal reasons have been excluded. Lastly, a conceptualisation which is appropriate to this realm has been proposed and recommendations

[14] Recital 51 of the GDPR: "Personal data which are particularly sensitive in relation to fundamental rights and freedoms merit specific protection as the context of their processing could create significant risks to the fundamental rights and freedoms".

[15] Notably, non-personal data, i.e., information not referred to an identified or identifiable individual, fall outside the scope of data protection legislation (art. 2 of the GDPR; art. 2 of Regulation 2018/1725).

[16] Germany, France, the Netherlands, and Italy have been selected for their institutional publications on AI governance. The United Kingdom has been excluded due to its withdrawal from the European Union.

have been drafted accordingly. Sources from academic literature in AI ethics have been reviewed to provide the necessary background to the institutional outcomes. Therefore, given the social impact of the public decision-making at stake, our analysis in conducted both at theoretical and institutional levels.

Security addresses issues regarding external threats to the systems (Vedder 2019b) used to perform data analysis. In particular, security should be ensured against "data access, modification and deletion, commensurate with its sensitivity" and the "risk of compromise of intended functions arising from both passive threats and active attacks" (Clarke 2019).

At the institutional level, it has been noted how the reliability of the infrastructure can be hindered by unauthorised third parties (HLEG 2019, p. 17). The relevance of security has also been considered in light of "by-design" principles and solutions (European Commission 2018, p. 9; BMWi 2018, p. 37), also in the design stage of training datasets (AGID 2018, p. 30). Finally, security has been proposed as a competitive asset to generate trust (BMWi 2018, p. 8).

In food safety risk assessment, security pertains to a wide range of areas, including the technical protection of data warehouses, confidential private data, food consumption and background information (*data* security). At the same time, it may also regard the technical robustness of AI systems and their capability to avoid fallacies, to be implemented in light of the *by-design* approach (*AI* security).

For what concerns data security, confidentiality, integrity and availability of the datasets (CIA) constitute an established design framework emerging from the literature (Olivier 2002). In our domain, confidentiality is necessary to limit access to private and personal (food consumption and background) information. Notably, the rationales underlying the confidentiality of these data are different: while the protection of commercially sensitive information fosters the legitimate interests of private parties, the secrecy of personal data is necessary for reasons linked to fundamental rights to privacy, data protection, and trust (Shin 2021a). However, the confidentiality of the datasets can be intended as a unified design requirement against illicit data processing and exploitation. As regards integrity, security should also be intended as the conceptual basis for safeguard measures that ensure that analysed data have not been altered. 'Data poisoning' can pose serious threats to the trustworthiness of the whole risk assessment procedures, including the use of MLTs.[17] Consequently, when safeguards are necessary to ensure confidentiality, their design should allow legitimate uses of data for the validation of assessment results.

Discussing AI security *strictu* sensu, it can be observed that, in light of the lower degree of autonomy of machine learning systems in our domain in comparison to others (e.g., self-driving cars or robots), the relevance of technical robustness is quite limited. However, the dimension of error handling (European Commission 2018) shall not be overlooked, in particular when designing MLTs systems used to support human decisions with large-scale effects (BMWi 2018, p.37). These points call for a discussion regarding the accountability mechanisms for the use of AI systems.

Accountability correlates with several concerns related to the use of Big Data and MLTs. It is well studied in theoretical frameworks (Pagallo 2017; Veale et al. 2018) and applied sectors (Lagioia and Contissa 2020), also from ethical perspectives (AI4People 2018). Taken together, these studies have noted how the identification of the entity responsible for the ethical and legal implications of the use of MLTs remain unclear (Shin 2021a). In turn, literature in food safety has critically described the progressive erosion of the barrier separating risk management decisions from the scientific risk assessment (Busuioc and Ambrus 2014). While EFSA's opinion become substantially binding (Kanska 2004), they are left without judicial scrutiny by European courts (Alemanno and Gabbi 2016, p. 41).

In the institutional debate, accountability mechanisms are needed to generate trust in the public (BMWi 2018, p.16). As noted (SIGAI 2019), accountability revolves around three stages of the deployment of AI systems: in the design stage, it concerns with the deployment of AI systems that are replicable (HLEG 2019, p. 17), especially to prevent improper uses, discriminations (BMWi 2018, p. 38) or violations of fundamental rights (European Commission 2018); in the monitoring stage, it allows auditing and verifiability of the system (Villani et al. 2018, p. 113); in the redress stage, which begins when the harm has occurred, accountability frameworks are needed to correctly allocate liability for the damage according to the existing regulations (Villani et al. 2018, p. 114) or from a moral perspective (SIGAI 2019, p. 5). Finally, accountability is crucial when public bodies make use of AI systems (AGID 2018, p.40).

In food safety risk assessment, the monitoring and the redress stage seem quite critical. On the one hand, EFSA competences and financial resources might not suffice to ensure the monitoring of algorithms used by third parties when submitting scientific evidence. State-of-the-art practices in risk assessment reveal the likelihood of scenarios in which EFSA scrutiny is performed with the support of MLTs (e.g., automated literature review) over MLTs-generated data

[17] Monsanto, one of the leading companies in plant protection products, has been accused of influencing scientific studies (McHenry 2018). As noted in the aftermath of the EU Fitness Check, the 'Monsanto Papers' scandal has significantly undermined the level of trust in the system.

(e.g., results in the reviewed literature).[18] Without an effective monitoring over submitted algorithms, the question pertaining to "cascade" accountability for MLTs-supported findings that influenced EFSA scientific opinions would remain unsolved. On the other hand, given EFSA's immunity from the jurisdiction of the EU Court of Justice for its scientific outputs(Alemanno and Gabbi 2016, p. 41), no entity would be liable for the 'probabilistic failures' that derive from high-quality data and working software. Therefore, it is necessary to correctly identify monitoring and redress measures that differ from the traditional framework for being focused on data quality (including the training of the dataset for supervised learning, which require a high degree of expertise), statistical errors (including precision and recall scores of the generated model) and human oversight.

Accountability principle also applies to the personal data processing activities at stake as a middle-out interface between top-down and bottom-up forms of regulation (*ex multis*, Pagallo et al. 2019). It might be useful to restate that statistical and research results are often aggregated, thus revolving around the scope of data protection laws for being related to groups rather than individuals. However, the proximity of processed data to individual and group identity and the consequences of data processing activities for random groups entail the need of taking into account the context in which data processing occurs, even when this falls outside the strict scope of data protection law (Vedder and Naudts 2017). This facet of the accountability principle has to be complemented by the interpretation of Fairness principle discussed below.

Fairness has been subject to a twofold interpretation in the literature. On the one hand, substantial fairness implies that the use of machine learning techniques should be steered towards the minimisation of the negative consequences of existing discriminations and the accessibility of the benefits due to AI (Jobin et al. 2019). Data-driven forms of discrimination are particularly problematic due to the three reasons highlighted by Hacker (Hacker 2018), i.e., their falling outside of the material scope of anti-discrimination law, the higher chances of indirect discrimination, and the placement of the burden to prove the discriminatory nature of the algorithm placed on discriminated individuals. In our scenario, individual discrimination is unlikely: data aggregation prevents the attribution of consumption patterns to a specific individual and EFSA's decision have no direct effects on individuals, let alone the inapplicability of data protection law. If *direct* cannot provide satisfactory results, *indirect* discrimination can be seen as a potential threat. It has been noted (Durante 2019, p.252) that only few elements, namely racial or ethnic origin can trigger the applicability of anti-discrimination legislation (e.g., in the EU, Council Directive 2000/43/EC), hence its inadequacy to confront other forms of discriminations, including the ones that originate on different, yet sensitive, grounds. On the other hand, as a corollary of substantial fairness, *procedural* fairness is meant to ensure that data analysis is performed by ensuring compliance with the law or ethical principles to prevent harm (Malgieri 2020). Possible technical solutions to guarantee procedural fairness include metrics (Wachter et al. 2021) and prejudicial remover techniques (Kamishima et al. 2012).

In the institutional AI charters landscape, the notion of fairness of MLTs mirrors the scientific debate. It has been referred to as the prevention of bias in results generated via MLTs (Villani et al. 2018, p. 121). Fairness is also linked to accountability measures (HLEG 2019, p. 7), in the sense that safeguards in place also guarantee effective redress mechanisms. Therefore, fairness entails inclusiveness and representativeness in training datasets (European Commission 2018, p. 13) to produce fair and balanced results, also by technical constraints (SIGAI 2019, p. 11). In general, the potential of MLTs should be oriented towards the minimisation of the adverse effects of existing differences, rather than allowing their increase (AGID 2018, p. 64) and towards the prevention of creating new wrongful differentiations (Vedder and Naudts 2017).

Given the kind of data used in food safety risk assessment, the most appropriate conceptualisation of substantial fairness seems to be a group-based approach that promotes equality among food patterns that aggregate individual preferences reflecting identified clusters (vegetarians, vegans, ethnic minorities). This is crucial for at least three reasons: first, food consumption data are proxy variables for attributes—such as ethnic origin or religious believes—that may originate subtle or even tacit discrimination if minority food patterns are not taken into account; second, discrimination might occur according to unconventional criteria, possibly linked to food preferences, such as veganism or vegetarianism, which are not yet referred to as grounds for discrimination in legislative systems, workplace rules, and other codes of conducts. Lastly, food and personal or social identity are inherently connected.[19] The existence of the individual association is true in biology, as we 'incorporate' substances into our body to stay alive, and in neuropsychology.[20] In addition, the correlation between social identity and food can be

[18] A more abstract version of this scenario is discussed in Vedder and Naudts (2017, p. 4).

[19] The quote 'Tell me what you eat, and I will tell you what you are' (Brillat-Savarin, 1841) well-summarises their connection.

[20] Research (Rozin et al. 1986) has shown that children progressively align their food preferences to adults only by growing and developing a more complete body.

found, for instance, in our occasional and often irreverent reference to other ethnic groups and people the *-eaters* suffix or expressions such as ‘Frogs’ for French, ‘Krauts’ for Germans, ‘Macaronis’ for Italians (Fischler 1988). In the light of this *substantial* interpretation of fairness, *procedural* measures can be drafted accordingly. These proposals might include techniques briefly mentioned above as well as novel measures that are context-specific for taking into account the information-types at stake. Dietary preferences change over time as we change our identity as individuals, groups or society. Fair grouping shall take into account the evolution of food consumption trends over time as individuals. Either way, procedural fairness measures should allow for measurement experiments already performed in other scenarios (e.g., in Shin 2020). As with other techniques, they also have to be discussed also in light of the Explainability principle (Zarsky 2016).

Explainability is a novel ethical principle primarily intended to promote algorithmic transparency and prevent opaqueness and "black box” MLTs systems. Explainability of MLTs is both an emerging research trend in AI ethics and AI in general, as well as a consolidated institutional and governance *desideratum*. It has been endorsed by the Commission as a key technical requirement (European Commission 2020) functional to the evaluation of fairness (European Commission 2018). The German approach (BMWi 2018, p.16) has identified its proximity to trust. This intuition was also confirmed by theoretical and practical research (Shin 2020; Shin 2021a, b). Explainable machine learning models allow the assessment of legal compliance (Doshi-Velez et al. 2017; Mittelstadt et al. 2019). Various proposals regarding the properties and the contents of explanations, both for the purposes of the GDPR (Sovrano et al. 2019) and other contexts (Miller 2019), have been made. The corollary of Explicability has been proposed by AI4People group (AI4People 2018) and the HLEG (HLEG 2019), which has referred to this principle as the capability of AI systems to communicate their operations and provide for a rationale for their output (intelligibility) and identify the responsible entity (“accountability”).

On the one hand, a considerable amount of research—including the eXplainable AI (XAI), research trend—is devoted to ensure that available machine learning techniques guarantee a sufficient understanding of their internal structure (global explanation); on the other hand, some have argued that users shall be able to interact with the system to grasp the "why and how" the decision has been taken (local explanation), including by means of intelligible user interfaces (Sovrano et al. 2019).

Explainability should be primarily conceived as the necessity to adopt models that provide for intelligible results. This is crucial to assess the validity of scientific assessments supporting the decision-making process.[21] Explainability should be guaranteed when individual and ensemble (i.e., combined) algorithms are in use (Gillespie 2014) or when MLTs-generated evidence rely on MLTs results whose logic is unknown to the second user (Vedder and Naudts 2017, p. 4), i.e., the risk assessor.

One EFSA commissioned study reported explainability scores for some of the scrutinised algorithms (IZSTO et al. 2017, p. 172). These findings are crucial to strike a balance between the degree of efficiency of MLTs and the certainty of their results. Hence, an assessment on explainability should be recommended to identify those scenarios in which the transition from traditional methods to MLTs would not be desirable for ethical reasons. The major implication of this finding is that, when scientific assessment is performed by using opaque MLTs, a comparable deterministic method shall be used to cross-validate the results.

Moreover, when unexplainable machine learning models prevent a deep scrutiny of possible risks due to the inscrutability of their working, the precautionary principle still applies for the risk managers.[22] In other words, the opacity of machine learning models shall be considered an integral part of the scientific uncertainty that might lead to the "most cautious decision" to be taken by risk managers in accordance with the precautionary principle. Such conceptualisation of scientific uncertainty explains the correlation between explainability and precaution. Contextualising algorithmic opaqueness in terms of scientific uncertainty—yet, within risk analysis procedures—can also fruitfully ease the attribution of accountability by providing a guideline for the cases in which risk managers neglected the results of algorithmic scrutiny.

Consistently with findings from the institutional documents (Villani et al. 2018, p. 115), research on explainability of MLTs is also needed in this domain. In particular, research on the perception and the social acceptability of MLT-generated results shall also focus on the role of explanations in public decision-making. Some argue that explanations should be user-centric and targeted to the receiver (Cowls et al. 2019). The major implication of this contribution in our domain is that there are at least three potential receivers of MLTs’ outputs: first, the risk assessor observes

[21] Differently from other scenarios in which the algorithm plays the role of decision-maker, neither EFSA competences nor risk assessment itself allow for automatic decision-making.

[22] The precautionary principle has been defined as the necessity that, "in cases of serious or irreversible threats to the health of humans or ecosystems, acknowledged scientific uncertainty should not be used as a reason to postpone preventive measures" (Jasanoff 2016). In data-driven risk assessment, the principle is usually understood as the case in which the lack of data originates scientific uncertainty towards a potential threat.

the evidence generated by the algorithm; then, the risk manager takes decisions on the basis of the explanations of the results provided by the assessor; finally, the general public perceives the decision taken by the risk manager. This calls for further research on design requirements of explanations and their effectiveness on receivers (scientists, decision-makers, citizens), perhaps with empirical verification similar to other domains (Shin 2020; Shin 2021a, b). The connection between Explainability and Transparency requires further investigation.

Transparency is commonly understood by the literature on MTLs within the taxonomical and conceptual framework of explainability (Rosenfeld and Richardson 2019) and related to the degree of "scrutinisability" of MLTs. Its relationship with fairness, accountability, and trust has been explored (Shin 2021a). Transparency is conceived as a design requirement capable of ensuring operational checks and audits of MLTs (Jobin et al. 2019; Cowls et al. 2019). This approach focuses on the social acceptability of the decisions taken by automated means by highlighting the role of the perception by the public. From these premises, transparency affects the public's perception of the decision in a manner to provide sufficient legitimacy. When this is the case, some authors suggest that transparency shall focus on the rationale underlying the decision rather than the process through which the decision has been taken, consistently with the principle of explicability (de Fine Licht and de Fine Licht 2020, p. 919).

The institutional debate on algorithmic transparency is highly influenced by scientific research on this topic. As it emerges from this debate, it relates to the ability of individuals to scrutinise the logic and the criteria underlying certain decision-making processes that make use of Big Data and MLTs (BMWi 2018, p. 38), thus overlapping with explainability (SIGAI 2019, p. 13). The notion has been used in a holistic manner to comprise data, algorithms and business models (HLEG 2019, p. 18) as a necessary corollary of auditability (Villani et al. 2018, p. 15). Importance has been given to the relationship between algorithmic transparency and the accountability of public bodies (AGID 2018, p. 11).

In our domain, transparency has been a consolidated principle since the first draft of food safety legislation. Authors have identified the scope of transparency in the disclosure of data held by EFSA (Conte-Salinas and Wallau 2016). In the *Hautala*[23] case, the European Court of Justice has stated that the publication of scientific data is crucial to an open discussion—especially in the case of scientific uncertainty—and ultimately fosters trust in EU institutions. These findings from literature and jurisprudence seem consistent with the new provisions of the Transparency Regulation. Despite the large consensus on transparency, it is not an unlimited principle, as it is constrained by the need of preserving confidential and personal data consistently with security measures. In the *Arysta* case,[24] the same Court has also stated that transparency of risk assessment regards the foreseeability of health-related effects of regulated products by the general public. This criterion is a viable way to contextualise the discussion on algorithmic transparency within the food safety framework. When MLTs are used to make predictions on future health effects, they could be subject to a certain degree of disclosure to the public. This would allow independent researchers (e.g., NGOs) to cross-validate the predictions. This interpretation of transparency is not in contrast with explainability requirements, but it should be understood as an ancillary principle: by giving access to algorithms and their results, independent reviewers can also validate them (e.g., counterfactually).

While the impact of the Transparency Regulation has still to be evaluated, some preliminary considerations can be provided on the need to make available data and algorithms in a way that allows public scrutiny. This might be cumbersome, as only a few people possess necessary skills and resources to perform independent evaluations of the results. Nonetheless, these reviewers should be empowered to carry out their research by having access to information other than'raw' data (e.g., metadata, source codes, log files, performance scores), as well as to have the possibility of communicating their findings in the case of scientific uncertainty.

6 Privacy as a meta-principle: from safety to p-safety

This subsection aims to cast light on what kind of privacy shall be discussed and how it relates to the SAFETY principles as a whole and with respect to each of them. Academic literature on privacy and AI is extensive. Published articles range from theoretical contributions (AI4People 2018; Cowls et al. 2019) to practical applications with ethical implications (e.g., in robotics (Ishii 2019), healthcare (Bartoletti 2019), recommender systems (Zhang et al. 2014), etc.). In the institutional debate, privacy is conceived as the protection of the private sphere (AGID 2018), both at individual and group level (Villani et al. 2018). When considered in its constitutional dimension in Germany, its role as a fundamental right is remarked (BMWi 2018). With specific regard to MLTs, due attention is given to the generation of

[23] Case T-329/17 Heidi Hautala and Others v European Food Safety Authority [2019] ECLI:EU:T:2019:142, 60.

[24] Case T-725/15 Arysta LifeScience Netherlands BV, formerly Chemtura Netherlands BV, v EFSA [2018] ECLI:EU:T:2018:977, 41.

training datasets that do not reflect biases or generate inaccuracies (European Commission 2018).

Despite its unanimous consensus in institutional charters (Fjeld et al. 2020, p. 21), the principles of privacy and data protection have been so far excluded from the SAFETY model. This is due to the technical and legal reasons discussed in this section. Privacy and data protection implications are a primary concern of AI institutional policy documents due to the pervasiveness of AI systems in individual private life, at mental, decisional, physical and informational level (Floridi 2013, Ch. 12.2). By discussing each of them in light of our technical and legal domains, evidence suggests that an outright adoption of this principle would be wrong.

First, mental and decisional privacy are not threatened by the use of MLTs in risk assessment due to the lack of any direct intervention on individuals' mind or decisions, including those who are surveyed to gather food consumption data. Second, the physical intrusion in the private sphere seems quite limited, especially in comparison to other technologies covered by institutional documents (e.g., facial recognition or AI-supported Internet of Things devices). Quite different would have been the case had real-time food consumption data collection systems been the state-of-the art.[25] Third, while data protection might have a strong relevance in data processing activities, technical and legal evidence discourages its explicit endorsement. On the one hand, food patterns are usually aggregated or otherwise anonymised, thus limiting the scope of data protection laws over-processed data; on the other hand, derogations of data protection laws for statistical and research purposes consistently limit the remit of rights granted to protect 'raw' personal information.[26]

Therefore, we will adopt a more cautious approach and consider (only) informational privacy as a meta-principle (P-) that enables all the SAFETY-based recommendations to respect the observation and the analysis of behaviours (food consumption) and characteristics (background information) linked to individuals and groups (Floridi 2017). For these reasons, an outright adoption of Privacy as a standalone principle would be wrong and likely to raise methodological concerns. Therefore, the SAFETY model could adopt informational privacy as a meta-principle, i.e., a principle primarily intended to serve and integrate the SAFETY model. Such meta- formalisation of informational privacy does not entail that less importance should be given to it in comparison to other principles. Instead, this conceptualisation simply clarifies that privacy is included in the model "in relation to" (P-) to other principles rather than as an independent component.[27]

The connection between Privacy and Security has already been introduced when discussing the implications of data processing activities with regards to individuals' food consumption and background information. We noted that, as these data might be used as proxy variable for sensitive inferences, data breaches and unlawful exploitations pose a serious threat to informational privacy. Such approach seems appropriate in light of the nature of the inferential potential of data at stake, hence the need of prioritising data protection measures for the personal information stored and analysed in this context. The same holds true for "AI security", i.e., the minimisation of threats posed to individuals by AI systems. In light of the limited degree of autonomy of MLTs in the domain at stake, Privacy relates with this facet of Security when considering that the use of MLTs might generate inferences that support decision-making processes with potential discriminatory issues.

Privacy can be a guidance to evaluate Accountability for the individual, collective and social implications of personal data processing. This can be done by two sets of measures: on the one hand, privacy-oriented accountability *strictu* sensu requires that compliance efforts are made and demonstrated (De Hert 2017) They include technical and organisational measures (*data protection by design*), staff training and records of data protection efforts. On the other hand, *latu* sensu accountability requires a risk-based aptitude towards the collective and social effects of large-scale data processing. While such formalisation of accountability may exceed the scope of data protection law, risks such as the possible re-identification or de-anonymization of surveyed individuals or data-driven biases call data controllers to take responsibility for their data processing activities.

[25] It has to be noted that, following the publication of 'EFSA Strategy 2020', the Authority has progressively endorsed the involvement of citizens through collaborative platforms and data crowdsourcing, both to foster trust in the Authority and to gather nearly real-time data. For instance, a 2017 EFSA tender (OC/EFSA/AMU/2017/02) asked participants to provide a "prototype design of a mobile app" to collect information pertaining to infants' consumption data, alongside parents' personal data regarding age, sex, region, and other information. To date, however, these data collection methods have not been fully deployed.

[26] Recital 156 GDPR: 'Member States should be authorised to provide, under specific conditions and subject to appropriate safeguards for data subjects, specifications and derogations with regard to the information requirements and rights to rectification, to erasure, to be forgotten, to restriction of processing, to data portability, and to object when processing personal data for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes'.

[27] A similar example of a meta-principle in the privacy domain can be found in the literature (Tamò-Larrieux et al 2021). In their discussion, authors argue that accountability 'serves as a meta-principle' in data protection law as it is provided by the legislation to fulfil the goals of other data protection principles by demonstrating compliance to their requirements. In our scenario, as it will be argued below, their relationship is reversed due to the limited applicability of data protection law.

Likewise, the conceptualisation of Fairness could benefit from the Privacy-oriented consideration that food consumption data analysis can discriminate on grounds that consist of sensitive attributes of individuals. This might have implications regarding *indirect* discrimination—the only possible in the context at stake—and group-level privacy (Floridi 2017). At the individual level, when conceptualised as a data protection principle,[28] Fairness requires that personal data is processed in a way that does not harm data subjects. This might trigger measures that prevent the unlawful exploitation of their data for purposes other than risk assessment, limited access to data, and, from a broader perspective, that other data protection law principles are respected (Malgieri 2020). Among them, data accuracy and data minimisation require careful balance exercises. On the one hand, data minimisation requires that the least possible amount of data is collected, whereas risk assessors might face the necessity of feeding algorithms with sufficient information to generate reliable results. On the other hand, hand, in light of the limited remit of data subjects' rights, data accuracy requires an effort to be made to adapt recorded food consumption data to individual preferences over time.

As discussed above, Privacy and Explainability principles are strictly related as a growing body of the literature is discussing explainable automated decision-making processes that bear a significant impact on individuals. However, this is not the case in our domain. Nonetheless, explanations of MLT-generated scientific results can be relevant for groups that result from the aggregation of food consumption data and mirror behaviours such as ethical or religious believing. This peculiar conceptualisation of the relationship between Privacy and Explainability stresses the relevance of the collective implications of inferences generated by MLTs rather than focusing on individual automated decision-making (Kuner et al. 2018).

The relationship between Privacy and Transparency might be cumbersome to grasp. Transparency is a key principle of data protection law[29] as it contributes to the informational self-determination of the data subject by making her or him aware of the nature of the processing, the entities involved and so on, ultimately fostering her or his free choice. However, the meaning attributed to Transparency in the SAFETY model—in essence, the widest possible availability of data and contextual information to replicate studies—differ from the one conferred in the context of data protection law. Therefore, when linked to the model, Privacy shall act a constraint to Transparency in giving access to raw personal information to third parties. This is necessary to prevent unlawful exploitations of data, including possible breaches of the purpose limitation principle.[30]

7 Implications and limitations

The aim of this research was to draft a set of principle-based recommendations for a trustworthy deployment of Big Data analysis through MLTs in the food safety risk assessment activities carried out at an EU-level. Such guidance seems to be needed as we enter in a transition stage in which risk assessment enters its 'datafication' age and risks and opportunities have to be evaluated. In particular, by contextualising ethical charters published by EU institutions and Member States to the technical and legal domains of risk assessment, this research has identified a P-SAFETY—Security, Accountability, Fairness, Explainability, Transparency, Privacy—model that has been used to draft high-level recommendations for a reliable use of Big Data and MLTs in the domain at stake. For technical and legal reasons, Privacy has been framed as a meta-principle that interacts with the others serving as an additional justification or as a constraint.

Let us first discuss some implications of this work. First, it contributes to an ongoing debate on the role of principle-based AI development, with a sort of case study relatively unknown to the ongoing debate on AI ethics, governance, and law. Nonetheless, it might be the case that scholars active in the fields in food technologies will gain interest towards this debate and will contribute by providing their valuable perspectives. As a consequence of the entry into force of the Transparency Regulation, the debate on food data and algorithmic governance could be further enhanced by such cross-disciplinary contributions.

Then, collective and social implications of food safety risk assessment are noteworthy. This work could raise attention towards the ongoing 'datafication' of risk evaluations and steer their future developments, Taken together, the adoption of these principles underlines the resilience of existing governance models (e.g., FATE) endorsed by European policymakers. Thanks to the ongoing debate on these principles-based models, their adaptation to a relatively unknown domain (that is, food safety risk assessment) has been proven possible. This case study shows that Security, Accountability, Fairness, Explainability, Transparency, and Privacy, are applicable to unknown research fields, at least in the form of a purely theoretical speculation.

Limitations of this study mirror its positive implications. In particular, this research solely provides a theoretical background without empirical verification. Surveys similar to

[28] Article 5(1) of the GDPR; Article 4(1) Regulation 2018/1725.

[29] Article 5(1) of the GDPR; Article 4(1) Regulation 2018/1725.

[30] Article 4(1)(b) of the GDPR; Article 4(1)(b) of Regulation 2018/1725.

the ones mentioned throughout the paper (e.g., Shin 2020, 2021a) seem necessary to corroborate our finding empirically. Then, food safety risk assessment does not constitute the only example of risk assessment: pharmaceuticals and chemicals present similar traits (e.g., relevant societal impact, risk analysis mechanisms, ongoing deployment of MLTs, conflicting data ownership issues, and so forth) and might constitute directions for further research. In particular, they can serve as a benchmark to verify the extent to which the proposed P-SAFETY model is extendable to other domains. Finally, the territorial scope of this analysis is limited and other jurisdictions (e.g. China) deserve attention for the ongoing globalisation of food markets. Notably, the Chinese approach to AI governance is a notable case study (Roberts et al. 2021) and can serve as an additional benchmark for the proposed P-SAFETY model.

The next question seems obvious: assuming that the P-SAFETY model can be implemented, how such regulatory framework could be drafted into concrete policy recommendations? A possible perspective could be the one proposed by the middle-out approach (Pagallo et al. 2019) due to its combination of bottom-up and top-down viewpoints, hence 'hard law' and 'soft law' (e.g., codes of conduct) instruments. In particular, this seems convenient due to the presence of multiple regulatory systems, different sets of rules, and equivalent design choices that could be implemented. Moreover, the foreseeable adoption of two important pieces of legislation in the EU—the Data Governance Act[31] and the Artificial Intelligence Act[32]—calls for an in-depth discussion on how these pieces of legislation, if approved, would be implemented in food safety risk assessment and neighbour domains.

Funding Open access funding provided by Alma Mater Studiorum - Università di Bologna within the CRUI-CARE Agreement.

References

AGID, Agenzia per L'Italia Digitale (2018) Libro Bianco sull'Intelligenza Artificiale al servizio del cittadino. https://ia.italia.it/assets/librobianco.pdf. Accessed 11 July 2021

AI4People (2018) AI4People | Atomium. https://www.eismd.eu/ai4people/. Accessed 15 Sept 2021

Alemanno A, Gabbi S (2016) Foundations of EU food law and policy: Ten years of the European food safety authority. Routledge, London

Barocas S, Selbst AD (2016) Big data's disparate impact. Calif l Rev 104:671

Bartoletti I (2019) AI in healthcare: ethical and privacy challenges. In: Riaño D, Wilk S, ten Teije A (eds) Artificial intelligence in medicine. AIME 2019. Lecture notes in computer science, vol 11526. Springer, Cham. https://doi.org/10.1007/978-3-030-21642-9_2

BMWi, German Federal Ministry for Economic Affairs and Energy (2018) Artificial Intelligence Strategy. https://www.bmwi.de/Redaktion/EN/Pressemitteilungen/2018/20180718-key-points-for-federal-government-strategy- on-artificial -intelligence.html. Accessed 12 July 2021

Brillat-Savarin JA (1841) Physiologie du gouˆt. Charpentier, Paris

Busuioc M, Ambrus M (2014) Blurred areas of responsibility: European agencies' scientific 'opinions' under scrutiny. The Role of Experts in International and European Decisionmaking Processes, p 383

Cavalli E, Gilsenan M, Van Doren J, Grahek-Ogden D, Richardson J, Abbinante F, Cascio C, Devalier P, Brun N, Linkov I, Marchal K, Meek B, Pagliari C, Pasquetto I, Pirolli P, Sloman S, Tossounidis L, Waigmann E, Schünemann H, Verhagen H (2019) Managing evidence in food safety and nutrition. EFSA J 17(S1):e170704, 17 pp. https://doi.org/10.2903/j.efsa.2019.e170704

Clarke R (2019) Principles and business processes for responsible AI. Comput Law Secur Rev 35(4):410–422

Conte-Salinas N, Wallau R (2016) The concepts of transparency and openness in European Food Law. In: Steier G, Patel K (eds) International food law and policy. Springer, Berlin

Cowls J, King T, Taddeo M, Floridi L (2019) Designing AI for social good: Seven essential factors. Available at SSRN 3388669.

de Fine Licht K, de Fine Licht J (2020) Artificial intelligence, transparency, and public decision-making: Why explanations are key when trying to produce perceived legitimacy. AI & Soc 35(4):917–926

De Hert P (2017) Data protection as bundles of principles, general rights, concrete subjective rights and rules: Piercing the veil of stability surrounding the principles of data protection. Eur Data Prot l Rev 3:160

Dubuisson C (2017) The French dietary survey on the general population (INCA3). EFSA Support Publ 14:12

Durante M (2019) Potere computazionale: L'impatto delle ICT su diritto, società, sapere. Mimesis, Milan

Dutch National Institute Public Health (2018) National dietary survey in 2012–2016 on the general population aged 1–79 years in the Netherlands. EFSA Support Pub 15(9):1488E

European Food Safety Authority (2014) Guidance on the EU Menu methodology. EFSA J 12(12):3944, 77 pp. https://doi.org/10.2903/j.efsa.2014.3944

EFSA (2015a) The EFSA data warehouse access rules. EFSA Support Publ 12(2):1–18

EFSA (2015b) The food classification and description system FoodEx 2 (revision 2). EFSA Suppor Publ 12(5):1–90

[31] Proposal for a Regulation of the European Parliament and of the Council on European data governance (Data Governance Act) COM/2020/767 Final.

[32] Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts COM(2021) 206 Final.

EFSA (European Food Safety Authority), Dujardin B, Bocca V (2019) Scientific Report on the cumulative dietary exposure assessment of pesticides that have chronic effects on the thyroid using SAS®software. EFSA J 17(9):5763, 49 pp. https://doi.org/10.2903/j.efsa.2019.5763

EFSA (2020) Commission White Paper on Artificial Intelligence—a European approach to excellence and trust. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf . Accessed 12 July 2021

European Commission (2018) Communication Artificial Intelligence for Europe. https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe. Accessed 12 July 2021.

Ferrario A, Loi M, Viganò E (2019) In AI we trust Incrementally: a Multi-layer model of trust to analyze Human-Artificial intelligence interactions. Philos Technol 33(3):1–17

Fischler C (1988) Food, self and identity. Information (international Social Science Council) 27(2):275–292

Fjeld J, Achten N, Hilligoss H, Nagy A, Srikumar M (2020) Principled artificial intelligence: mapping consensus in ethical and rights-based approaches to principles for AI (January 15, 2020). Berkman Klein Center Research Publication No. 2020-1. https://doi.org/10.2139/ssrn.3518482

Floridi L (2013) The ethics of information. Oxford University Press, Oxford

Floridi L (2017) Group privacy: a defence and an interpretation. In: Taylor L, Floridi L, Van der Sloot B (eds) Group privacy. Springer, Berlin

Floridi L (2018) Soft ethics and the governance of the digital. Philos Technol 31(1):1–8

Floridi L, Taddeo M (2016) What is data ethics? Philos Trans R Soc A 374:2083

Floridi L, Cowls J, King TC, Taddeo M (2020) How to design AI for social good: seven essential factors. Sci Eng Ethics 26(3):1771–1796

Gillespie T (2014) The relevance of algorithms. In: Media Technologies: Essays on Communication, Materiality, and Society: The MIT Press. Retrieved 15 Sept 2021. https://mitpress.universitypressscholarship.com/view/10.7551/mitpress/9780262525374.001.0001/upso-9780262525374-chapter-9

Gilsenan MB (2015) Data handling: observatories/databases/data storage/legal framework: EFSA data collection. In: Options Méditerranéennes. Series A: Mediterranean Seminars. CIHEAM-IAMZ, Zaragoza (Spain)-EFSA, European Food Safety Authority, Paarma, Italy

Hagendorff T (2020) The ethics of AI ethics–an evaluation of guidelines. Mind Mach 30:99–120

HLEG, High Level Expert Group on AI (2019) Ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai. Accessed 15 Sept 2021

Hwang Y (2020) The role of affordance in the experience of blockchain: the effects of security, privacy and traceability on affective affordance. Online Inf Rev 44(4):913–932

Ishii K (2019) Comparative legal study on privacy and personal data protection for robots equipped with artificial intelligence: looking at functional and technological aspects. AI & Soc 34(3):509–533

Istituto Zooprofilattico Sperimentale del Piemonte (2017) Liguria e Valle D'Aosta; Unità di Biostatistica, Epidemiologia e Sanità Pubblica del Dipartimento di Scienze Cardiologiche, Toraciche e Vascolari dell'Università degli Studi di Padova; Dipartimento di Scienze Cliniche e Biologiche dell'Università degli Studi di Torino; Zeta Research s.r.l., Trieste, 2017. EFSA supporting publication 2017:EN-1254. 311 pp. https://doi.org/10.2903/sp.efsa.2017.EN-1254

Jasanoff S (2016) The ethics of invention: technology and the human future. WW Norton & Company

Jaspers S, De Troyer E, Aerts M (2018) Machine learning techniques for the automation of literature reviews and systematic reviews in EFSA. EFSA supporting publication 15(6):EN-1427. 83 pp. https://doi.org/10.2903/sp.efsa.2018.EN-1427

Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics guidelines. Nat Mach Intell 1(9):389–399

Kamishima T et al. (2012) Considerations on fairness-aware data mining. In: 2012 IEEE 12th International Conference on Data Mining Workshops. IEEE, pp 378–385

Kanska K (2004) Wolves in the clothing of sheep? The case of the European Food Safety Authority. Eur Law Rev 5:711–727

Kazutoshi S (2018) You are what you eat: a social media study of food identity. In: arXiv preprint arXiv:1808.08428

Kocharov A (2009) Data ownership and access rights in the European Food Safety Authority. Eur Food Feed Law Rev 4(5):335–346

Kuner C, Cate FH, Lynskey O, Loideain NN, Millard C, Svantesson DJB (2018) Expanding the artificial intelligence-data protection debate. Int Data Privacy Law 8(4):289–292

Lagioia F, Contissa G (2020) The strange case of Dr. Watson: liability implications of AI evidence-based decision support systems in health care. Eur J Leg Stud 12:245

Lazarou C, Karaolis M, Matalas A, Panagiotakos DB (2012) Dietary patterns analysis using data mining method. An application to data from the CYKIDS study. Comput Methods Progr Biomed 108(2):706–714

Lodge J (2003) Transparency and EU governance: balancing openness with security. J Contemp Eur Stud 11(1):95–117

Lynskey O (2014) Deconstructing data protection: the 'added-value' of a right to data protection in the EU legal order. Int Comp Law Q 63(3):569–597

Malgieri G (2020) The concept of fairness in the GDPR: a linguistic and contextual interpretation. In: Proceedings of the 2020 Conference on fairness, accountability, and transparency, pp 154–166

Malgieri G, Comandé G (2017a) Why a right to legibility of automated decision-making exists in the general data protection regulation. Int Data Privacy Law 7(4):243–265

Malgieri G, Comandé G (2017b) Sensitive-by-distance: quasi-health data in the algorithmic era. Inf Commun Technol Law 26(3):229–249

Marcos SV, Rubio MJ, Sanchidrián FR, de Robledo D (2016) Spanish National dietary survey in adults, elderly and pregnant women. EFSA Support Pub 13(6). https://doi.org/10.2903/sp.efsa.2016.EN-1053

McHenry LB (2018) The Monsanto Papers: poisoning the scientific well. Intl J Risk Saf Med 29(3–4):193–205

Miller T (2019) Explanation in artificial intelligence: Insights from the social sciences. Artif Intell 267:1–38

Mittelstadt B (2019) AI ethics–too principled to fail? arXiv preprint arXiv:1906.06668

Mittelstadt B, Allo P, Taddeo M, Wachter S, Floridi L (2016) The ethics of algorithms: mapping the debate. Big Data Soc 3(2):1–21

Morley J, Floridi L, Kinsey L, Elhalal A (2020a) From what to how: an initial review of publicly available ai ethics tools, methods and research to translate principles into practices. Sci Eng Ethics 26(4):2141–2168

Morley J, Machado CCV, Burr C, Cowls J, Joshi I, Taddeo M, Floridi L (2020b) The ethics of AI in health care: a mapping review. Soc Sci Med 260:113172. https://doi.org/10.1016/j.socscimed.2020.113172

Naydenova S, de Luca L, Yamadjako S (2019) Envisioning the expertise of the future. EFSA J 17(S1):e170621. https://doi.org/10.2903/j.efsa.2019.e170721

Olivier MS (2002) Database privacy: balancing confidentiality, integrity and availability. ACM SIGKDD Explor Newsl 4(2):20–27

O'Neil C (2016) Weapons of math destruction: how big data increases inequality and threatens democracy. Crown, New Yosk. ISBN: 9780553418828

Pagallo U (2017) From automation to autonomous systems: a legal phenomenology with problems of accountability. In: 26th International Joint Conference on Artificial Intelligence, IJCAI 2017. International Joint Conferences on Artificial Intelligence, pp. 17–23

Pagallo U, Casanovas P, Madelin R (2019) The middle-out approach: assessing models of legal governance in data protection, artificial intelligence, and the Web of Data. Theory Pract Leg 7(1):1–25

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

Roberts H, Cowls J, Morley J, Taddeo M, Wang V, Floridi L (2021) The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation. AI & Soc 36(1):59–77

Rosenfeld A, Richardson A (2019) Explainability in human–agent systems. Auton Agent Multi-Agent Syst 33(6):673–705

Rozin P, Hammer L, Oster H, Horowitz T, Marmora V (1986) The child's conception of food: differentiation of categories of rejected substances in the 16 months to 5 year age range. Appetite 7(2):141–151

Sapienza S, Palmirani M (2018) Emerging data governance issues in big data applications for food safety. In: Kő A, Francesconi E (eds) Electronic government and the information systems perspective. EGOVIS 2018. Lecture notes in computer science, vol 11032. Springer, Cham. https://doi.org/10.1007/978-3-319-98349-3_17

Sette S (2011) The third Italian national food consumption survey, INRAN-SCAI 2005–06–part 1: nutrient intakes in Italy. Nutr Metab Cardiovasc Dis 21(12):922–932

Shin D (2020) User perceptions of algorithmic decisions in the personalized AI system: perceptual evaluation of fairness, accountability, transparency, and explainability. J Broadcast Electron Media 64(4):541–565

Shin D (2021a) The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. Int J Hum-Comput Stud 146:102551

Shin D (2021b) Embodying algorithms, enactive artificial intelligence and the extended cognition: you can see as much as you know about algorithm. J Inf Sci 1:1–14

SIGAI, Special Interest Group of AI (2019) Dutch AI Manifesto. http://ii.tudelft.nl/bnvki/ wp- content/ uploads/ 2019 / 09 / Dutch- AI-Manifesto- 2019 . pdf. Accessed 22 July 2020.

Simpson C (2016) Data protection under food law post: in the aftermath of the novel foods regulation. Eur Food Feed Law Rev 11(4):309–314

Sovrano F, Vitali F, Palmirani M (2019) The difference between explainable and explaining: requirements and challenges under the GDPR. In XAILA@ JURIX

Taddeo M, Floridi L (2018) How AI can be a force for good. Science 361(6404):751–752

Tamò-Larrieux A (2018) Mapping the privacy rationales. In: Tamò-Larrieux A (ed) Designing for privacy and its legal framework. Springer, Cham, pp 27–43

Tamò-Larrieux A, Mayer S, Zihlmann Z (2021) Not hardcoding but softcoding privacy. Technol Regul (**forthcoming**)

Van der Meulen BM (2013) The structure of European food law. Laws 2(2):69–98

Veale M, Van Kleek M, Binns R (2018) Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In: Proceedings of the 2018 Chi Conference on human factors in computing systems, pp 1–14)

Vedder A (2019a) Mind the gap. Managing the expectations of legal scholars turning to ethics for help where the law does not yet provide answers. In: Centre for IT and IP Law (ed) Rethinking IT and IP law. Intersentia, Cambridge, pp 305–312

Vedder A (2019b) Safety, security and ethics. In: Vedder A, Schroers J, Ducuing C, Valcke P (eds) Security and law. Legal and ethical aspects of public security, cyber security and critical infrastructure security. Intersentia, Cambridge, pp 11–26

Vedder A, Naudts L (2017) Accountability for the use of algorithms in a big data environment. Int Rev Law Comput Technol 31(2):206–224

Villani C, Bonnet Y, Rondepierre B (2018) For a meaningful artificial intelligence: towards a French and European strategy. Conseil national du numérique

Wachter S, Mittelstadt B, Floridi L (2017) Why a right to explanation of automated decision-making does not exist in the general data protection regulation. Int Data Privacy Law 7(2):76–99

Wachter S, Mittelstadt B, Russell C (2021) Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law. West Virginia Law Rev (**Forthcoming**)

Yakovlev PA, Walter P, Guessford WP (2013) Alcohol consumption and political ideology: what's party got to do with it? J Wine Econ 8(3):335–354

Zarsky T (2016) The trouble with algorithmic decisions: an analytic road map to examine efficiency and fairness in automated and opaque decision making. Sci Technol Human Values 41(1):118–132

Zhang B, Wang N, Jin H (2014) Privacy concerns in online recommender systems: influences of control and user data input. In: 10th Symposium On Usable Privacy and Security ({SOUPS} 2014), pp 159–173