Regulation & Governance

Regulation & Governance (2021) doi:10.1111/rego.12417

Certification systems for machine learning: Lessons from sustainability

Kira J.M. Matus

Division of Public Policy, Hong Kong University of Science and Technology, Hong Kong, China

Michael Veale

Faculty of Laws, University College London, London, UK

Abstract

Concerns around machine learning's societal impacts have led to proposals to certify some systems. While prominent governance efforts to date center around networking standards bodies such as the Institute of Electrical and Electronics Engineers (IEEE), we argue that machine learning certification should build on structures from the sustainability domain. Policy challenges of machine learning and sustainability share significant structural similarities, including difficult to observe credence properties, such as data collection characteristics or carbon emissions from model training, and value chain concerns, including core-periphery inequalities, networks of labor, and fragmented and modular value creation. While networking-style standards typically draw their adoption and enforcement from functional needs to conform to enable network participation, machine learning, despite its digital nature, does not benefit from this dynamic. We therefore apply research on certification systems in sustainability, particularly of commodities, to generate lessons across both areas, informing emerging proposals such as the EU's AI Act.

Keywords: artificial intelligence, certification system, governance, machine learning, sustainability certification.

1. Introduction

Machine learning systems, often colloquially called "algorithms" or even "artificial intelligence" systems, are a type of software distinguished by the way that they "learn from experience." It is said that a machine "learns" if, after being exposed to data, its subjectively measured performance at a given task improves (Mitchell 1997).[1] They have been at the receiving end of a range of recent critiques concerning varied social impacts accompanying their commercial use. These critiques have generally centered upon their potential to create and perpetuate unwanted discrimination, both within applications (Custers *et al.* 2013; Barocas & Selbst 2016) and more structurally (Gangadharan & Niklas 2019) and the difficulties apparent when attempting to scrutinize these systems in practice (Hildebrandt & Gutwirth 2008; Edwards & Veale 2017). One way this furore has so far manifested itself in the policy sphere has been through a range of national and international reports and initiatives targeted squarely at machine learning systems (see e.g. House of Commons 2018; House of Lords 2018; Committee of Experts on Internet Intermediaries (MSI-NET) (2017); Craglia *et al.* 2018) – some of which have been branded as "ethics-washing" (see generally Bietti 2020). Another has been the high-profile unearthing of obscure and unlitigated but relevant strands of pre-existing law, such as the "automated decision making" provisions present in European data protection law in various forms since their genesis in 1970s France (Bygrave 2001; Edwards & Veale 2017). Few previous topics in technology law and policy have created as much buzz as the governance of machine learning.

In this paper, we argue that the nature of machine learning issues, and the range of different actors that become involved in this process of learning from experience in a global context, allows us to usefully characterize machine learning through lenses more familiar to the study of socioenvironmental issues and sustainability more

Correspondence: Michael Veale, University College London, London, UK. Email: m.veale@ucl.ac.uk

* Conflicts of interest: The authors are not aware of any conflicts of interest.
Accepted for publication 27 May 2021.

broadly in commodities and supply chains. This characterization facilitates the bringing together of the two fields, particularly with the aim of examining the nature of challenges that those seeking to regulate machine learning using private governance regimes may face.

We proceed as follows. In Section 2, we first note that there are key similarities between policy-relevant characteristics of commodities and machine learning systems. Certain aspects of both can be classified as *credence qualities* – qualities that cannot be assessed, even after use (such as the impacts that the production of a particular food stuff, such as coffee, had on the environment, compared to the fundamental rights issues in collecting the data to build a model). Furthermore, both machine learning and commodities commonly analyzed in relation to their socioenvironmental characteristics can be seen through the lens of the *value chains* involved in their production and use. We examine similarities in three areas: (i) networks of labor used in the creation of machine learning algorithms; (ii) the core-periphery inequality in the economics of system design; and (iii) the fragmentation and modularity of value in machine learning systems.

In the sustainability of commodities, these characteristics, likely due to their fit, have coincided the use of *certification* as a governance tool. Certification has also been prominently explored in the area of machine learning. In Section 3, we illustrate how these two views of certification are surprisingly far apart. We attribute this largely to the focus of discussion around machine learning certification centering on networking-style standards, an area many technologists in that area have experience with, and we argue that these foundations are inappropriate and likely to be ineffective. We instead move to see what lessons research and practice around sustainability certification can provide to machine learning, applying research findings about the former to the sociotechnical features of the latter.

Given the broad and general-purpose nature of machine learning systems, and the fast-emerging nature of business models and practices, we rely throughout the paper on a broad array of sources to identify and highlight issues and comparators at a relatively early stage. In seeking to do so, we draw on a range of scholarship from many disciplines, high-quality journalism, and policy documents. This is particularly important in a domain where trade secrets and NDAs typically restrict a thorough external analysis of machine learning products or practices within a particular sector or company.

2. Value chains illuminate underexplored governance issues in machine learning

Prevailing academic and policy analysis of machine learning has typically viewed it as a service provided by a software object. This service might output *prima facie* discriminatory results (Barocas & Selbst 2016), be insecure to cybersecurity attacks (Huang *et al.* 2011; Veale *et al.* 2018b), or be opaque and difficult to scrutinize (Edwards & Veale 2017). When analyzed in its deployment context, it might, for example, create feedback effects in interaction with its environment, such as contributing to over-policing of already marginalized communities (Lum & Isaac 2016; Ensign *et al.* 2018), or engender organizational pressures to transfer it to new and unsuitable contexts (Veale *et al.* 2018a).

All these issues can, and are, analyzed only using information about the software object, and information about its deployment context and downstream effect. But what happens upstream to create this deployed system? To do so requires analysis not of the functionality of the service of machine learning systems, but analyzing machine learning as a *good*.

To understand what is omitted, we now introduce some concepts relating to the qualities of products from the governance of value chains and commodities (Darby and Karni 1973; McCluskey and Loureiro 2005). *Experiential* qualities are those that are possible to observe upon an item's use – consider a banana's taste or longevity. For a machine learning model, experiential qualities might include its explicability to certain audiences, or its performance offline, on test set data. Experiential qualities contrast with *credence* qualities, which due to their largely invisible or unattributable natures, are generally a matter of buyer or user trust or belief. Credence qualities can be further distinguished. Some credence qualities are difficult or costly to *observe*, largely because they do not leave a mark on the product. In commodities, these may be labor or environmental standards in production; in machine learning, they may also include aspects such as privacy or unethical experimentation in the collection of training data. Other credence qualities are costly to *attribute*. In commodities, these notably include health impacts from consumption, or environmental impacts from disposal. In machine learning, we may place aspects of deployed performance in this category: systemic biases, feedback loops, or insecurity to attacks in practice that may be (methodologically) difficult to detect before or even during deployment, but ultimately result from upstream production or design choices. Impacts on human rights through processes of data collection or labeling (Roberts 2019), or impact on emissions from large model training (Bender *et al.* 2021), are further examples of credence attributes in machine learning. We do not attempt to present an exhaustive list here. Issues such as these may not be immediately evident without costly, dedicated study, and may require additional data to evaluate (Veale & Binns 2017). It is important to clarify what this terminology is not intended to denote: upstream harms may certainly be experiential in the sense that they consciously experienced by individuals and communities (Petty *et al.* 2018).

These three categories are mapped in Table 1.

To go backwards before an item's purchase or use to a point where credence qualities could be observed, the lens of the *value chain* is particularly useful. Value chains are a way of understanding production and consumption patterns in an interconnected, global economy. In the commodity sector, value chains span from the producers of raw materials, such as agricultural smallholders, to end consumers that may be located in a very different part of the world. These end users might be consuming a product that has been significantly transformed through its combination with other converging value chains, processing, and the provision of services to producers all over the world. Value chain analysis highlights the links between globalization and existing and emergent forms of inequity, and calls attention to how systems might be reshaped to mean that the creation of new, higher value goods increases the living standards of those across the component parts of their production systems.

The lens of value chains has emerged as particularly useful in understanding systems where any single actor lacks much unilateral leverage over the production processes in a sector. This includes governmental actors, who in some accounts are forcibly sidelined by globalization (Cutler *et al.* 1999) or in others step aside to legitimate further economic integration (Sassen 1996). Single states have little power to manage processes that span multiple jurisdictions, and state-sponsored international actors, organizations, and efforts have also had limited buy-in and impact. Disempowered actors can also include firms in these fragmented processes who are positioned in such a way that does not afford them significant sectoral control. Yet at other times, the analysis of global value chains can highlight how specific actors maintain control, even despite a lack of explicit vertical integration.

Commodities such as coffee and tea can be easily seen through this lens (Neilson & Prichard 2009). Value chain analysis also shines light on composite products such as electronics or clothing, which often integrate many components and transnational production processes through complex subcontracting processes that can be opaque even to the core actors involved.

Value chain analysis is commonly used to consider issues of sustainability, given the position of environmental and labor characteristics as difficult-to-observe credence attributes (see e.g. Mol & Oosterveer 2015). We argue this usefully applied to issues of justice in machine learning models.

Why would a frame developed for tangible commodities and composite products be applicable to intangible, informational ones? Machine learning systems might be portrayed in the media and in popular culture as "smart" software and traditionally might be seen as more akin to intellectual property than a physical commodity (such

Table 1 Attributes of commodities versus attributes of machine learning

	Commodities	Machine learning
Experiential *difficult to observe until use*	Quality (taste, durability, etc.)	For example, measurable performance; interpretability; energy use from querying
Credence (origin) *difficult to observe*	Production conditions and footprint	For example, training process (exploitation, privacy, energy use, or emissions from training)
Credence (use) *difficult to attribute*	Post-consumption impacts (e.g. health, disposal effects)	For example, systemic biases, robustness, and resilience

as subject to a European database right), but a closer look at their development and maintenance emphasizes significant parallels which are often overlooked. Here, we draw upon several of the characteristics of value chains highlighted in the research literature (see generally Gereffi *et al.* 2005) to make the case that such a framing can provide insights into the development of machine learning systems, too. These are:

- *Networks of labor* used in the creation of machine learning systems;
- Clear *core-periphery inequality* in the economies of system design; and
- The *fragmented and modular nature of value* in machine learning systems.

We now address each in turn.

2.1. Networks of labor

The vast majority of deployed machine learning systems are built on datasets collected from the physical and social worlds.[2] Many of these datasets are created by either *observing* individuals or from individuals *explicitly contributing* to them.

Observed data might be recorded unbeknownst to the individuals generating it.[3] For example, systems that help understand the meaning of words by locating them as a point in high-dimensional space relative to other words – word vectors – are broadly trained on huge corpora of text data, such as the entirety of English Wikipedia, or from web crawls (Mikolov *et al.* 2013; Pennington *et al.* 2014). They might also be taken from individualized surveillance the user is unaware of, such as the data captured by Facebook Pixel, a covert web tracker present on a large proportion of commonly used sites on the Web (Karaj *et al.* 2018), or through trackers embedded in mobile apps (Binns *et al.* 2018). Even when individuals are unaware of these activities, populations are set up as the sources of these extractive activities by the design of the infrastructure and business models that surround them (Zuboff 2015). These activities that leave a data shadow – content creation, browsing, indexing, and the like – can be construed as labor undertaken by individuals around the globe (Morozov 2019). Indeed, the information systems and software people use every day are and have for many years been constructed to elicit such labor from individuals by re-organizing their activities (Agre 1994; Gürses & van Hoboken 2018, pp. 594–95). Software products are today developed using "agile" methods where the users are constant subjects of experimentation in real time. This "agile" approach integrates this constant labor of re-organization into the development processes of much software today (Gürses & van Hoboken 2018). Following from the way labor has been commodified, and increasingly informationalized, as an industrial input, contemporary practices of data extraction and processing themselves create new sources of "raw materials" framed as inputs to particular types of productive activity (Cohen 2019). Despite usually not being able to "own" data that result from these processes under current intellectual property regimes, either in the United States or the EU, companies act as legal entrepreneurs to piece together an array of mechanisms from terms-of-service contracts, NDAs, trade secrets and even privacy law to jealously guard access to and enclose the infrastructures creating and processing these "raw materials" (Cohen 2019).

Other forms of labor for machine learning involve significantly more *explicit collection practices*. Supervised machine learning, the most common type deployed today, requires *labeled* data – historical data that have included the value(s) a system should be able to predict when given a new data record. For example, a social media post might have a label rating it for abuse; a new support ticket for a helpdesk might have the most relevant FAQ, or the length it took to resolve; a section of CCTV footage from a store might have label indicating whether shoplifting is visible within it. Machine learning practitioners sometimes call this a "ground truth." There are many forms labor to create labels can take. Web users will likely be familiar with Google's *reCAPTCHA* product, which requires individuals to answer questions before they are allowed to proceed with an activity such as accessing a site (such as "which of the following photos contains a picture of a shop front"). *reCAPTCHA* has a dual purpose design: (i) creating an effective barrier against automated clients attempting to undermine the integrity or availability of online services through methods such as registering fake accounts or contributing to distributed denial of service (DDoS) attacks and (ii) to create an extensive pool of labeled data for training machine learning models by asking questions that such models currently struggle to answer reliably.

Online projects from the "Scalable Cooperation" lab at MIT Media Lab indicate possible future directions for such explicit labor. In the "Deep Empathy" project,[4] individuals are asked to choose from two photos that "*make [them] feel more empathic [sic] toward victims of Syria crisis [sic]*" to "*train [their] algorithm to get better*." In the "Moral Machines" project, individuals are presented with "moral dilemmas, where a driverless car must choose the lesser of two evils, such as killing two passengers or five pedestrians."[5] Effectively, both these studies ask humans to do one thing only humans can do: emotional labor. As machines get better at comparatively objective tasks such as assessing shop-fronts, it may be that emotional laboring is one of the main areas left to present to users to assess their humanity, as well as train the emotional analysis or "affective computing" (Picard 1997) systems commonly deployed in areas such as marketing today. Indeed, CAPTCHA exercises deigned to interface with aspects of individuals' emotions are already being developed by computer scientists (Seering *et al.* 2019).

Even more explicit labor for machine learning exists. Labeled datasets are commonly created using both informal "crowdworkers" who work from home and are paid by the job through platforms such as Amazon Mechanical Turk, or through contractors and sub-contractors, particularly of large technology firms. They often labor to moderate content posted to social media platforms for abuse or "toxicity," or to label illegality such as child pornography, incitement to violence, or content intended to radicalize individuals (Roberts 2019). A range of investigative news stories have also uncovered poor conditions, limited psychological support, tenuous employment, and even evidence that moderators themselves can be radicalized as a result of the content they are judging (Chen 2014; Newton 2019). These workers are too often treated like "just an API call" by both technology firms and machine learning researchers who use them to build labeled datasets (Silberman *et al.* 2010). This sits closely alongside many other forms of labor and material extraction used in production of machine learning systems (Crawford & Joler 2018; Ensmenger 2018).

2.2. Core-periphery inequality

Geographic disparities are prominent in the creation if machine learning systems. Cheaper markets, such as the Philippines, are often used to outsource the labeling tasks discussed in the previous section. Many of these tasks are potentially distressing or invasive and are undertaken for a fraction of the wage of countries in the Global North (Chen 2014).

Modern software development and machine learning in particular rely on experimentation (Gürses and Van Hoboken 2018; Bird *et al.* 2016). Experimentation in the production of information has a history of core-periphery issues we should be wary of machine learning replicating as the sector grows. For the last 30 years, the pharmaceutical sector has been increasingly locating clinical trials in developing countries (Petryna 2009, p. 13), where costs can be one tenth of those in the United States (Glickman, Cairns, and Schulman 2009, p. 816). The availability of willing participants (often in need of basic medical treatment) is considerably higher, important to easily reach the statistical power needed to trial drugs treating rare conditions or conferring only slight benefits (Weigmann 2015) – even though these drugs are rarely relevant to the health issues in the country hosting the trial (Glickman, Cairns, and Schulman 2009).

We see this trend toward identifying and isolating willing sites for participation in machine learning particularly in urban environments. *Sidewalk Labs* was a project by Google's parent company, Alphabet. The company was developing a "smart city" using the testbed of the Quayside, a 12 acre plot of land in downtown Toronto. With strong backing and permissions provided by the city, a critical and recurring question in the media surrounded the extent to which a wide array of sensors and experiments were going to be run on Toronto residents in order to provide training data for systems to be sold and deployed elsewhere (Sauter 2018). To the concern of many, the CEO of Alphabet Inc and co-founder of Google, Larry Page, proposed in 2013 that to avoid the issues of regulation, it might be possible to "set aside a part of the world" for permissionless innovation around connected technologies (Ingraham 2013). The distribution of benefits relative to experimentation and surveillance between "testbed" cities and beneficiary cities is yet to be seen, although the Sidewalk Labs project was canceled in May 2020, ostensibly due to uncertainty from the economic fallout of COVID-19.

Insofar as knowledge about human behavior and institutions is transferable, but requires large-scale experimentation in order to crystallize into machine-learned models, it is likely that such experimentation will take place where the economic and regulatory calculus facilitates it, to the benefit of where its deployment is most valuable.

2.3. Fragmented and modular value creation

Deployed machine learning systems are often misleadingly portrayed in the media (and at times, in the scholarship) as if they are an instantiation of weak "artificial general intelligence" – a single, polymathic machine capable of achieving a wide array of tasks. This notion stems at least in part from how the research institutes receiving the most publicity, including Google DeepMind and Open.AI, have this as their stated goal. In practice, however, this is not how machine learning generally works, nor how it is shaping up in business contexts. Machine learning systems must still be built separately for specific tasks, such as image recognition or speech synthesis, often specialized further for the type of data in a particular application domain.

A range of technology companies seek to create and monetize the "best" machine learning technologies for particular, general applications. There are two main business through which this currently occurs and is increasingly seen (Veale *et al.* 2018b). The first is through the *licensing of application programming interfaces (APIs).* This approach sees a firm, usually with unrivalled access to datasets, train a machine learning model, and make it available through a black-box, pay-per-query interface, usually through a cloud service. Currently, such models commonly do things such as speech recognition and synthesis, affective (emotion-based) computing, image recognition, or translation – question-answer tasks amenable to query over the Internet. The second is through *packaged model trading*, where a trained model is given to users (potentially in a protected format) to be run locally. This move is often used where models have to be queried rapidly with low latency or in areas of patchy or no connectivity. To this end, companies such as Apple and Google have been building hardware and software packages to more efficiently run complex trained models, such as *CoreML* or *TensorFlow Lite*. It is also used when querying or augmenting models in the cloud would involve the transfer or potential leakage of sensitive data, for example in medical contexts.

These models themselves are unlikely to be singular, but bespoke to different customers and specialized by sector. Tools such as Google *AutoML* are designed to allow companies to augment these models with their own datasets in order to specialize them further. These models in turn are integrated, alongside other software and potentially more machine learning models, into a final product. The suppliers of them might be swapped (compare for example *Google Cloud Vision API*, *Microsoft Azure Face API* and *Amazon Rekognition*) or the datasets used to augment the models changed or added to. These systems in turn might be integrated into other services or APIs, which are sold to customers downstream. This significant interweaving of data sources makes analysis of the value chains created by these networks of data and analytics both necessary and challenging.

3. Proposed private governance of machine learning

Many of the key features of both credence issues and value chains make them a challenge from a governance perspective. Voluntary programs are a popular approach to governance in the area of commodity value chains, commonly turned to in an effort to coax complex systems toward credence goods and social ends such as environmental sustainability or heightened labor rights. Such a turn has many names in the literature, including private regulation or governance, private law, or private authority regimes (Cutler *et al.* 1999; Mattli & Büthe 2003; Cashore *et al.* 2004).

Private governance has emerged as an important part of discussions around ensuring sustainability. In part, this is because sustainability is a shared global challenge, characterized by trans-border effects and common pool resources, which have made co-operation between states challenging. Information technologists are often familiar with private governance mechanisms too, given the role organizations such as the Internet Engineering Taskforce (IETF) or the World Wide Web Consortium (W3C) have played in the governance of the Internet and the Web respectively (see generally Harcourt *et al.* 2020).

A range of policy tools exist to attempt to tackle social issues in globe-spanning supply and value chains, which might be successfully adapted to machine learning (Matus 2010a). These include:

- Standards, certification and labeling;

- Databases, tools and information sharing;
- Awards and recognition;
- Economic incentives;
- Market creation and preferential purchasing;
- Technology mandates;
- Regulatory restrictions;
- Research funding;
- Partnerships and collaborations;
- Technical assistance.

In many cases, such tools are used in concert to achieve some policy end. Here, we primarily consider the first category – standards, certification, and labeling.

Standards, certification, and labeling are one policy tool that has already seen activity in the area of machine learning and society. Conceptually, *certification systems* contain three elements: a *standard*, a *certification*, and a *label.* A standard lists "specifications and/or criteria for the manufacture, use, and/or attributes of a product, process, or service." Certification is the "process, often performed by a third party, of verifying that a product, process, or service adheres to a given set of standards and/or criteria." Lastly, labeling is the "method of providing information on the attributes, often unobservable, for a product, process or service" (Matus 2010b). While for some products, the label is synonymous with a brand identity that is used to communicate to household level consumers, the label can also be thought of as a straight-forward way for certified entities to signal the qualities (especially credence qualities) of a product. For the purposes of this study, labeling covers any method of clearly communicating that a product has been certified (formally evaluated) to comply with a particular standard. Labels communicate authenticity and specific (usually unobservable) qualities, which is crucial for the success of these private governance systems.

We can contrast this three-step approach to proposed standards for machine learning. The efforts for standards around machine learning that have been developed come predominantly from a very particular institutional heritage of voluntary standard-setting organizations with norms and structures originating in the years before and after the start of the twentieth century (Yates & Murphy 2019).

At the time of writing, there is significant activity occurring in such standard setting bodies concerning machine learning. The International Organization for Standardization (ISO) technical committee on artificial intelligence, ISO/IEC JTC 1/SC 42, has published three standards on "Big Data" focusing on vocabulary, implications for existing standards, and thoughts on the ways forwards for standards and is developing nine standards on artificial intelligence, including areas such as bias and trustworthiness. Some standards already exist from national ISO member bodies addressing ethics, robotics, and artificial intelligence. British Standard (BS) 8611:2016, for example, *Robots and Robotic Devices: Guide to the Ethical Design and Application for Robots and Robotic Systems*, attempts to build a categorization of harms from these systems (see generally Bryson & Winfield 2017). Additionally, 10 standards relating in varying ways to social dimensions of machine learning systems are currently under development, with one published, by the Institute of Electrical and Electronics Engineers Standards Association (IEEE-SA), a spin-out of its related professional association with long ties to official national standards bodies (Yates & Murphy 2019, p. 225).[6] These "IEEE P70xx" series standards focus on issues such as data management and algorithmic bias and are being developed by a range of stakeholders. IEEE P7001, for example, is a proposed transparency standard aiming to set out measurable and testable level or tiers of transparency, and a toolkit for self-assessing against these levels (Bryson & Winfield 2017, p. 119).

Predominantly, these standard-setting initiatives have so far been focused on achieving consensus on the terminology and analytical frameworks that should surround these emerging sociotechnical systems. Much of the potential influence they might have would be as a "meta-standard," laying out terminological use applicable to substantive standardizing processes.

While sustainability standards are conceived as a part of a proposed certification system, the institutions underpinning these proposed machine learning standards have typically treated consensus on terminology and information exchange as a major goal in and of itself, providing significant benefits. This is due to a category error between two uses of the term. Above, we have defined a standard as comprising *specifications and/or criteria* *for the manufacture*, *use*, *and/or attributes of a product*, *process*, *or service*. However, for institutions such as the ISO, IEEE or IETF, a standard is instead, more narrowly, *a structured protocol for conforming the activities of multiple nodes in a network* (Cohen 2019, p. 217), producing a networked governance arrangement where actors with infrastructural control can gain policy dominance without processes of auditing or assurance to ensure any specifications or criteria are being met. In practice, if you are not using IEEE 802.11, the common wireless computer networking standard behind, you will not be able to communicate with devices around you which are expecting you to use this standard – the well-known power of architectural rules in the information domain (Reidenberg 1998).

Sustainability certification systems, not being based on networking, have consensus as an instrumental rather than intrinsic aim. Convergence supports an efficient and workable system from the perspective of auditing or labeling, and it is hard to see how a hugely fragmented system could carry much legitimacy. Yet it does not bring enforcement. Machine learning is a product of data analysis, not a layer of a network (or a good where compatibility and interoperability are major end-goals). But the institutions currently considering and developing machine learning standards that claim to seek to govern credence properties, such as IEEE P70xx, either seek convergence for networked governance or produce conceptually clarifying meta-standards, which are not by themselves the substantive standards needed in a certification system. This does not seem like a very strong fit if a certification system is a governance arrangement expected or sought from these efforts.

In contrast, organizations designing standards for sustainability certification have to design them to fit governance arrangements that do not draw their power from network effects. The policy challenges posed by machine learning, particularly those that relate to its credence properties outlined in Section 2, may benefit from reaching agreed standards that improve, for example, energy efficiency or working conditions for labelers. However, unlike in networking, standardization cannot be both the desired and as well as the main mechanism of change. The similarities between commodities and machine learning outlined in Section 2 mean that if we are to envisage private governance systems relating to the credence properties of machine learning, it would be instructive to attempt to apply lessons from across the sustainability certification sector to this new potential regulatory endeavor. It is to this we now turn.

4. Learning from sustainability certification systems

There have been many attempts to effectively and responsibly regulate credence goods in relation to sustainability concerns. As actors across the supply chain have become concerned with the socio-environmental impacts of production and have struggled with the challenges of assessing the qualities of credence goods, one governance tool that has become popular is the use of third-party certification systems. One underlying logic is that by providing consumers information about the credence quality of these goods (i.e. how sustainably they were produced), they will be able to command a price premium, which will cover the added costs of both the more sustainable production methods, and the certification process. This price premium, in turn, will incentivize more producers to join the program, improving their production processes and contributing to an improvement in sustainability in the value chain (van der Heijden 2012; Komives & Jackson 2014).

Establishing a virtuous (and effective) cycle via sustainability certification has proven to be difficult in practice. Yet despite the challenges, a number of sustainability certification programs have managed to grow and become financially viable. Private voluntary sustainability standard systems are a market-based approach to promoting sustainable production and business practices. Adoption of these sustainability standards is intended to be voluntary: the standards are not created, run, or required by governments or government regulation (Cashore 2002). Instead, voluntary sustainability standard systems are non-government initiatives that seek to drive sustainable production and consumption by creating market demand for sustainable products, and a supply to meet that demand. They are designed to help buyers (both consumers and businesses) identify sustainably produced products, and they guide producers, forest managers, mine and tourism operators, and factory owners and others in the choice of sustainable practices (Komives & Jackson 2014). Many others have been at least partially successful, though frequently still struggle to move beyond a reliance on donor funding to support their operations (Barry *et al.* 2012). Despite the difficulties in developing and growing sustainability certifications, their number has grown dramatically. The Ecolabel Index currently tracks 463 ecolabels in 199 countries and 25 sectors (www.ecolabelindex.com) – though not all of these are full certification programs.

There are several features of this growth in sustainability certification that provides an important set of lessons that may be relevant to machine learning. The first is the set of features of a robust certification system.

As previously discussed, these systems consist of up to three parts, which are meant to cover the full set of functions of a regulatory system: standard setting, behavior modification, and information collection (Hood *et al.* 2001):

1. A set of *technical standards* (which can cover management systems, methods of production, outcomes, or some combination therein). These standards answer the question of *what* is being regulated (what behavior needs to be modified). The choice of technical standards reflects a *theory of change*, which links something that can be measured or monitored to the underlying problem.
2. A *certification process* (which includes monitoring and enforcement) to ensure that the standards are being met/followed. The certification process answers the question of *how* will it be regulated – specifically mechanisms for monitoring and enforcement, including appropriate auditors, who are responsible for information collection.
3. A *labeling program* (to provide information on the credence qualities to consumers). This answers the question of *communication* of participation to the "market."

In the sustainability domain, there is an ISO meta-standard that sets out best practices for the process of establishing a sustainability certification program (Lambin & Thorlakson 2018), similar to the meta-standards discussed above. There are also private meta-standard efforts. One example is that ISEAL Alliance, founded in 2002, which is a global membership body for sustainability certifications. To become a member, these certifications must be deemed "credible," by virtue of meeting ISEAL's Codes of Practice[7] (Lambin & Thorlakson 2018). Members also share best practice, work together to promote innovation, and provide an important platform for sharing and problem solving. The development of particular Codes of Practice, and associated guidelines, helps member certifications improve their effectiveness (Komives & Jackson 2014). While only a fraction of sustainability certifications are members (currently there are 19 full members and five associate members), they remain an important platform for the dissemination of best practices.

Experience in the sustainability space has underscored the underlying regulatory scholarship, which indicates that successful regulatory systems have multiple requirements. Lodge and Stirton (2010), in their expansion on the work of Hood *et al.* (2001), expand into five distinct areas of consideration:

1. The decisionmaking process that leads to the creation of a regulatory standard in the first place;
2. The existence of a regulatory standard for affected participants within the regulated policy domain;
3. The process through which information about the regulated activities is being gathered and how this information is "fed back" into standard-setting and behavior-modification;
4. The process through which regulatory standards are being enforced;
5. The activities of the regulated parties themselves.

The experience of sustainability certification systems in each of these areas is worth exploring, as doing so is likely to provide guidance to the analogous value chain system and governance of the credence properties of machine learning. In exploring this in the section below, we consider studies and observations that relate to a range of commodities. This is not to generalize findings across distinct commodities with their own challenge, but to focus on what can be transferred as potentially analytically useful frames and questions for further investigation in the distinct machine learning context.

4.1. Decisionmaking and standard-setting

In sustainability certification, multiple stakeholders have come together to address what has been seen as a gap in more traditional, government-based certification systems for particular commodities. Some of these gaps were the result of the cross-jurisdictional nature of global commodity value chains (similar to the case argued for machine learning). Organizations such as the ISO and the ISEAL Alliance have developed procedures for creating a governance structure and decisionmaking framework to enable a process of standard-setting. These multi-stakeholder processes differ from traditional regulatory systems, in that participants need to find ways to decide on which actor groups are represented in the process, and how to balance conflicting interests, without the government acting via its authority capacity (Cashore 2002; Bernstein & Cashore 2012; Bennett 2017). Different certification systems have faced criticisms when they get this balance "wrong" – for example when they are critiqued for being overly (or insufficiently) representative of industry interests, at the expense of producer, environment, or civil society groups (Cashore 2002; Auld & Gulbrandsen 2010; Bennett 2017). The tangible result of getting decision structures "right" is supposed to be reflected in the legitimacy of the standard in the market – although this translation is complex and may further depend on both the intended audience for a label (e.g. business-to-consumer or business-to-business) and the broader context.

Such legitimacy is related to the aims of the standard, and the mechanisms by which the standard seeks to achieve these aims. One reason that sustainability is a useful analogue to deployed machine learning is that it too is, at its core, a complex, adaptive system that is difficult for even the most experienced experts to completely understand in its sociotechnical entirety, not to mention predict. For sustainability, the ultimate concern is a set of outcomes (labor practices, environmental conditions, livelihoods) – but these are often difficult to measure, observe, and attribute to any particular set of activities. For many sustainability certifications, early standards focused on processes and procedures (i.e. farming practices) (Komives & Jackson 2014). These are more clearly observable than the direct and indirect impacts of farming activities. However, with any practice or process-based standard, there remain questions regarding whether the connections between practice or procedural standards and outcomes are significantly robust, and the degrees of stringency required to balance the need for uptake, but also behavioral change, to deliver impacts (Gulbrandsen 2004; Cashore *et al.* 2007; Auld *et al.* 2008; Barry *et al.* 2012). They also can present challenges when the "good practices" outlined in the standard are not linked to concrete processes for implementation, or when there are gaps between the content of the standard and actual practices. This results in challenges for transparency, and quality, of auditing and evaluation.

More recently, performance-based standards have been proposed, which seek to measure impacts directly and work backwards from there (Veale & Seixas 2015). Unlike the prescription of particular processes, these give firms flexibility in how to achieve their goals (Gunningham 1996), although they may lack the capacity or understanding of how to best meet them as a result. In sustainability, these standards might measure, for example, how much carbon dioxide particular production emits (following a permissible methodology), and seek its reduction as a result.

Performance-based standards in social impacts machine learning would be in line with recent work in "debiasing" machine learning or making it explicable in particular ways, such as that published by a growing field of computing scholars at conferences such as *ACM FAccT*, *NeurIPS*, and *ICML*. This work is particularly important as these conferences, heavily attended and organized and presented at by large technology firms, appear to have significantly influenced industry practice. A performance-based standard would make a specific definition of, say, fairness, in accordance with a particular philosophy (Binns 2018) and would seek to measure systems' performance against this metric. They contrast with either prescriptions for particular safeguards or technologies, or "red lines," such as recent calls for bans on facial recognition or granular behavioral advertising.

Yet a comparison with sustainability highlights a downside. Society is concerned with multiple impact categories in relation to both commodities and to the use of technology in the context of broader social systems. One thing that has emerged from the (admittedly limited) evaluations of sustainability certifications is that no program has been able to address the entire range of sustainability concerns effectively. For example, Fair Trade is strong on labor practices and livelihood issues, but quite weak in terms of environmental outcomes. Many of the other certifications, such as FSC and Rainforest Alliance, have set strong environmental standards, but are much weaker on social elements (Barry *et al.* 2012). At a narrow level, the practical outcome is twofold. One, that buyers often have to choose between desired outcomes when given choices of different certified products in the market. And second, that many producers are pressured to become certified by multiple programs, which can be costly and even contradictory (Abbott *et al.* 2017). For machine learning, the implication is that, for practical reasons, there may need to be different certifications (or policy approaches entirely) for labor practices and data use, for example – but the impacts on those who need to be certified, and the power of the market to "pull" entities by choosing certified products needs to be considered in any regulatory design.

More concerningly perhaps, certification systems also place a market mechanism at their core, and the impacts possible from sustainability certification of commodities are limited to those that are achievable within a market paradigm. Certification systems operate within existing economic relationships, international trade agreements, physical infrastructures, and regulatory environments, with limited potential to change any of those aspects of the system. Opting for market-based regulation further risks narrowing the space for dialogue away from policy interventions that may change these broader rules and conditions (Parker *et al.* 2017). Insofar as those characteristics are the main forces restricting change toward sustainability, certification has limited transformative potential.

Similarly, in machine learning, the quantification of deeply historically rooted issues such as discrimination to make it subject to a type of reduction inherently limits the issues in scope. Indeed, seeing it this way might "shy from an analysis of systematic forms of injustice," as "technologically mediated discrimination exists alongside other forms of discrimination that contribute to the systemic marginalization of individuals and groups marked by social difference" (Gangadharan & Niklas 2019). Reducing discrimination to that which can be measured and mitigated at scale automatically and computationally does not challenge existing infrastructures and systems and may even serve to legitimize and entrench them.

Furthermore, the credence properties that are amenable to standard-setting and certification in machine learning, and which resemble sustainability the most, such as labor, environmental, or privacy standards during training, only encompass a subset of the policy issues associated with machine learning. Systems that are environmentally friendly, free of bias according to a relevant definition, or trained in an ethical and responsible manner can still be put to reprehensible ends or incorporated into abusive business models (Keyes *et al.* 2019). In commodities, lifecycle assessments, for example, coffee, indicate about half of the environmental footprint happens at a stage controlled by a user (e.g. brewing, disposal), meaning that certification is inherently limited in the issues it can tackle even using this simple frame (Humbert *et al.* 2009).

If there *is* an acceptance that standards are a required approach, sustainability shows us that the question of *what kind of standard* is not inconsequential, and that there may be a trade-off between what it is possible to standardize, and the desired outcomes of the standard (Barry *et al.* 2012). More broadly, restricting the range of governance tools to only what is possible to standardize itself shuts down broader political questions which standard-setting organizations have never faced head-on (see e.g. on the IETF, Cath & Floridi 2017; on data ethics, Taylor & Dencik 2020, pp. 8–9). There is broad agreement that a world with more interoperable machinery creates new social and economic possibilities. There is also relatively strong consensus that international trade of commodities is not going anywhere fast, and that markets are here to stay. Yet both in sustainability and in machine learning, there is no consensus that relevant emerging and incumbent sociotechnical and socioeconomic structures should persist. Should countries in the Global South continue to be engage only in comparatively lower-value parts of the value chain that are mostly transformed into higher value products abroad, even if such production is, in isolation, subject to good labor conditions and environmental protections? Should individuals in socially precarious situations and marginalized communities – or simply individuals online – be more heavily surveilled and assessed for "risk" using an increasing array of automated systems, often powered by machine learning, even if those systems are "fair" and "transparent" according to an agreed upon standard? Realistically, we might expect a number of differently focused certification programs to emerge and "fight it out" in the market, but competition among different notions of labor standards or different notions of machine learning fairness will not be the same as debates around the broader rules of the game and the infrastructures that support them – particularly if such standards are supported by the very institutions, platforms, and infrastructures that those debates may wish to challenge. Certification systems may be legitimate governance tools for aspects of issues around machine learning – but they will not likely be well-suited to them all.

4.2. Information, feedback, and enforcement

Initially, many sustainability certification actors focused on the first two pieces of the regulatory structure: bringing stakeholders together and setting up governance mechanisms for standard-setting. But there are only the first two elements of a successful certification system. A strong program of assessment, monitoring, and enforcement is also required. In fact, these are often limiting factors for growth of certificates (Matus 2014). Ensuring compliance can be costly, and often relies on scarce capacity, in terms of the availability of appropriately skilled auditors (or similar "regulatory intermediaries") (Abbott *et al.* 2017) – which for legitimacy reasons are almost always organizations external to the standard-setting organization (Cashore 2002; Auld & Renckens 2017). There are also potential principal-agent problems, wherein the auditors are under competitive pressure with other auditing agencies. They may compete for human resources, or drive towards efficiency, which may not always align with stringency (Auld & Renckens 2017). Recent scholarship has broadened the scope of intermediaries to any actor that acts as a "go-between" in the relationship between a regulator and its target regulatees. Brès, Mena and Salles-Djelic (2019) develop a typology based on two dimensions: formal/informal and official/unofficial. This typology is useful to better understand the positive and negative impacts that different types of intermediaries can play in these regulatory systems. Work by Kourula *et al.* (2019) divides the work of intermediaries in these systems into four generic roles: creating and/or organizing, coordinating between programs, supporting the implementation of a program, and voicing an opinion.

Auditors would generally fall into the category of formal intermediators (Brès, 2019). Auditors can play a crucial role in successful implementation of standards. These bodies are usually the ones responsible for interpreting and translating standards into practice on the ground. They are the groups interacting most closely with producers, and their feedback to the certification programs is an important source of information for the certification programs. There is increasing attention being paid to their role in the effectiveness of regulation generally, including in voluntary certification systems (Abbott *et al.* 2017). In general, while regulatory intermediaries can bring benefits of expertise, capacity, and legitimacy to a certification program, they can also act in ways that are self-interested, or even corrupted (Abbott *et al.* 2017). The challenges associated with ensuring that the auditing system in place is robust, meets the needs of all parties, and has the required technical capabilities available are all elements that need to be considered in any effort to bring certification to machine learning.

The question of enforcement and monitoring is one that machine learning should take quite seriously. The current state of technical standard setting, for example the efforts by IEEE, resembles industry-led self-certification programs that emerged in the chemical industry in the late 1980s and early 1990s in the wake of the Bhopal tragedy. These efforts, known as *Responsible Care*, were designed to address public concerns about the potential health and safety impacts of chemical manufacturing on local communities. In an effort to forestall regulation, the chemical industry put together a program that compelled firms to undergo a series of planning exercises and self-certify these activities. Adherence to *Responsible Care* was made a requirement for membership in the American Chemical Council, although it spread through industry internationally. However, after much criticism over the years, the program had to turn to outside auditors and an overall more rigorous process of standard setting (including more community participation), monitoring, and enforcement. Even so, it has never really escaped its reputation as being a weak attempt by industry to avoid more serious regulatory efforts (Gunningham 1995; Gunningham *et al.* 2004; King & Lenox 2000; King, Lenox, and Terlaak 2005; Lenox 2007; Lenox 2006).

Insights on the importance of legitimate, formal intermediaries in private governance need to be taken seriously by the machine learning community. Similar to the chemical industry, actors outside the sector are unlikely to distinguish between different (firms) in the sector in the case of negative outcomes. Negligence or carelessness can reflect poorly on the whole system and reduce the social license to operate, leading to tougher, more intrusive regulatory systems. Especially if the sector is viewed as having used less legitimate self-regulation to deflect government regulation (Gunningham et al. 2004). Data-driven systems are no stranger to ineffective self-regulatory efforts. In privacy in particular, it has been a frequent feature, pushed especially in the United States and widely criticized for its inefficacy (Gellman & Dixon 2011). A proliferation of self-regulatory codes of conduct, public commitments, certificates, and privacy seals in this space has been conceptually complicated by the integration of these modes of governance as provisions in European data protection law (Bennett & Raab 2018), but this reification of them in statute has not generally been appraised as leading to their effectiveness.

The ineffective nature of these instruments was most apparent in the grisly demise of the EU-US Safe Harbor agreement, a data sharing agreement between the two blocs felled at the hands of the Court of Justice of the European Union in 2015 in *Schrems v. Data Protection Commissioner*. Its failure was linked to the considerable skepticism about how the private governance and certification seal instruments such as *TRUSTe* the agreement relied upon really improved privacy practices on-the-ground (Charlesworth 2000). These self-regulatory mechanisms were seen by the Court as weak in a range of ways. It was apparent that "a significant number of certified companies did not comply, or did not comply fully, with the safe harbor principles" (*Schrems*, para 21), and the Court noted that while certification might plausibly bind the certified actor, it did not give Europeans recourse against the US Government compelling a certified organization to hand over their personal data. They were replaced by a broadly similar system of self-certification under the EU-US Privacy Shield Framework, which was promptly also declared invalid for similar reasons in 2020 by the Court of Justice in *Facebook Ireland and Schrems*. Ongoing conceptual challenges continue to stress surveillance law in the context of machine learning (Kosta Forthcoming).

Certification as a mechanism now exists in the General Data Protection Regulation (GDPR), although a literal reading indicates it is only capable of certifying data controllers and processors against the standards in that law, without creating a presumption of compliance (see Leenes 2020), rather than capable of being applied as a certificate to machine learning systems themselves to demonstrate their credence properties. Indeed, in general the processing surrounding credence properties will likely be outside the GDPR's material, and potentially territorial, scope once it reaches a controller who wishes to use it. It is "inspired by the stages of the international ISO/IEC 17065," a meta-standard laying out appropriate processes for conformity assessment of the certification of goods and services, rather than specific to social issues around machine learning (Kamara & de Hert 2018, p. 17). Regulators do, however, have significant investigative and transparency powers when certification is being sought, which might serve to increase accountability and monitoring (Henderson 2017). Such a process is familiar in Internet and telecoms governance as a form of co-regulation, where a body with statutory regulatory authority monitors a code or certificate generally developed and applied by industry (Marsden 2011). However, while Internet co-regulation, as discussed by Marsden, can operate on the regulatory access points or bottlenecks of a network, such access points are not always so easy to identify for software, which can be copied, open-sourced, or run locally, on end-users' devices. Data protection regulation illustrates this in part, particularly when seen through a "decentered" framework (Yeung & Bygrave Forthcoming) Examining the regulation of commodity flows, along with the role of enforcement and intermediaries in the sustainability sphere, may help conceptually bridge Internet co-regulation theory and the governance of machine learning value chains.

Just as it is difficult to provide assured history of a particular coffee bean in the supply chain, it is difficult to keep track of datasets and changing systems. In sustainability certification, enforcement has also required the development of a broader set of certificates that guarantee that certified commodities have been identified and separated from non-certified and non-controlled commodities as they make their way along the supply chain, particularly where they pass through areas with weaker legal systems. These "chain of custody" certificates are, in many cases, a structural prerequisite to a functioning system with limited fraud and deception. In the field of data engineering, the efficient and effective design of a system to capture and manage meta-data about the production process of an object – whether a cash crop or a piece of data – is called *provenance* (Carata *et al.* 2014). Consequently, the functional requirements of systems to track physical objects are extremely similar to those that track data objects. While physical objects are not vulnerable to certain threats (e.g. a coffee bean cannot be covertly duplicated), they are also costly to attach provenance information to in a secure manner. This means that schemes are vulnerable to, for example, a quantity of experientially high-quality certified crop being swapped out for low quality produce that can be passed off as certified and sold for a higher value due to its credence attributes. While it might be expensive to "fingerprint" a coffee bean, more options are available to do this in relation to data, as indicated by the literature on creating data policies and metadata with "sticky" attributes during data flows (Miorandi *et al.* 2020). Recent work has explored applying cryptographic methods and the tools from the data provenance literature to machine learning systems (Kroll *et al.* 2017; Singh *et al.* 2019) – effectively focusing on providing code-based constraints, opposed to legal or organizational constraints alone (Reidenberg 1998). However, while technological tools can help, capturing and verifying these data remains challenging in environments with many actors, boundaries, and different systems (Singh *et al.* 2019; Miorandi *et al.* 2020), and the surrounding governance-related accountability pre-requisites required for a successful certification system that can be examined and enforced cannot be underestimated or reduced to technical considerations. The practical experience of chain-of-custody certification in sustainability schemes may help machine learning certification avoid practical and unexpected pitfalls in this regard.

4.3. Actions of the regulated (and non-regulated) entities

A certification system is only successful if it can incentivize behavioral change at sufficient scale.

One issue that has emerged in the sustainability arena is the "adoption paradox." This stems from the fact that in order to be financially viable, a standard requires a minimum level of uptake. To meet this, in some cases, the standard must be sufficiently low, lest adoption prove too difficult and costly, in which case, the low adoption rate makes it impossible for the program to remain viable. Yet set the bar too low, and the adopters are either those that are already performing at a high(er) level, or just below it, and who can easily improve. Once certified, there is little, if any, incentive to improve performance (Cashore 2002; Auld *et al.* 2008).

One potential strategy is then a program wherein the standards are raised over time, or tiered, which provides incentives for continuous improvement. Another implication is that certification is often a useful complement to a strong regulatory "floor," wherein traditional state regulation protects against the worst actors, and certification provides an incentive for best actors and overall improvement (Cashore *et al.* 2007; Bernstein & Cashore 2012). One problem that has emerged in the sustainability space is the lack of coordination between different certification systems – and between certified and non-certified production.

In the case of sustainability, there are a number of competing certification systems for particular commodities. This has a number of potential impacts: the first is direct competition for market share between standards (and a potential "race to the bottom"). Similarly, multiple certifications for a given commodity potentially allow producers to choose to align with those that require the fewest (and therefore least costly) adjustments to their current practices, in which case certifications end up maintaining the status quo, as opposed to catalyzing change. In some cases, producers may be under pressure to align with multiple certifications, yet this can be quite costly, and may involve conflicts between standards. In some cases, multiple certifications may be synergistic or reinforcing; in others, they may actually work against the goals of the systems (Lambin & Thorlakson 2018). This complicates attempts to understand effectiveness, by requiring a more systemic approach, both for those interested in evaluating these tools in practice, but also for those considering designing new systems.

Another consideration for the design of certification systems, in terms of their potential to catalyze change, is the question of positive spill-overs, wherein non-certified entities pick up practices that the observe being used by their neighbors, even if they do not undergo certification themselves (Borck & Coglianese 2009; Jaffee 2014). A closely related issue is essentially the reverse challenge – the impact that non-certified actors, especially those using "dirtier" (in the case of sustainability) practices may have on the larger system. For example, one uncertified firm upriver could essentially negate the benefits of downstream certified production.

For machine learning, a core question is whether a mosaic of certified (potentially via different programs) and non-certified actors/algorithms would undermine the goal(s) of certification or create a new set of issues that would need to be addressed via some other sort of intervention.

The fact that standards are being developed by the IEEE points to interesting comparison for this tension. The IEEE, like organizations such as the IETF, have arguably been successful in previous and quite arcane standards efforts, but these efforts are distinguished by being primarily a co-ordination challenge. This is not to say that technical standards have not been political – for example amongst telcos, who often preferred Internet standards that maintained their centralized power (such as X.25) over standards that were more centrifugal in nature (such as TCP/IP) (see Ryan 2011). Yet machine learning standards for social issues in general do not present coordination challenges in the same way – and do not present the coordination benefits that result from solving them, either. Company A pledging not to use exploitative methods to gather training data does not mean that the system they build will not technically play with that of Company B, downstream. This encourages us to cast a closer eye on the governance implications of a plurality – or a monoculture – of machine learning standards.

In particular, it is important to see these dynamics within the context of the large incumbent actors who act as primary funders of stakeholder meeting-places, such as the *Partnership on AI*. Primarily, these are the GAFAM – Google, Amazon, Facebook, Apple, and Microsoft, as well as large technology firms outside of North America such as Baidu and Tencent. It seems unlikely adoption of any standard can be seen as a success without the participation of one or more of these firms. Yet, insofar as compliance involves changes to a process or technology rather than an entire business model, these firms have extraordinarily high technical compliance capacity themselves. Effectively, all of these organizations each employ a significant proportion of the global research capacity in areas such as fairness, transparency, and privacy and as a result are likely to be able to ensure any new technologies they implement are compliant in ways that might be prohibitively costly for other, smaller players. As a result, it is questionable how much any standard designed by these firms would be suitable for smaller technology firms, and the impact on competition more generally. In sustainability certification, organizations such as Fair Trade have received criticism on issues of scale, such as around certifying entities such as plantations. Such criticism hinges on the basis of broader questions – structurally, can a plantation (as opposed to e.g. smallholder cooperatives) ever be 'fair'?

The issue of positive spillovers points to another very important lesson that has emerged from sustainability certification. Because of the system complexity, it has been very difficult to study the direct impacts of certification (reasons for this include the difficulty in developing proper counterfactuals, data collection challenges, and a focus on process as opposed to outcome in standards). But a set of "indirect" outcome pathways has emerged as important avenues of change (Barry *et al.* 2012; Bernstein & Cashore 2012). Among these are certification as:

1. Developing agreement and reducing conflict among key stakeholders;
2. Providing a niche for experimentation and development of new practices; and
3. Acting as a demonstration for practices that could be scaled up via more "traditional" policy and regulation.

Current thinking on certification sees these programs as just one of many in a toolbox; appropriate for particular contexts, as opposed to a panacea. Practitioners and policymakers have been working across many commodities to develop and share best practices, and the spill-overs – between certified and non-certified, between certification systems, and between certification systems and state-based regulations – are all important elements that have been recognized in the sustainability space that are of use to those considering the use of similar concepts in machine learning.

5. Conclusions

As the machine learning community considers the role that certification could play in governing connected social impacts, it should be mindful that the closest analogy to the systems they might wish to develop may not be found in the domains of networking, electronics or telecoms, or even Internet regulation, but in sustainability, environmental governance and policy.

Lessons from this space particularly concern establishing all the elements of robust certification system, which must include well-developed processes for standard-setting, certification monitoring and enforcement, and labeling/information provision. There are some key questions that need to be answered for the development of standard and certification programs in this space. *What will be subject to the standard? Is it the programmer? The algorithm? The data? Some combination? Does the standard (behavior modification) actually track to the outcome of concern? How will adherence to the standard be monitored, and by whom? Is there capacity for this kind of program? And what form of labeling or communication will be affected? What is the target audience for the label? Who is going to pay, and will the costs of such a program be worthwhile?*

More broadly, considering the parallels between sustainability and machine learning places emphasis on some of the broader credence concerns surrounding machine learning, which may be exacerbated or made more visible as it matures. As we have argued, machine learning can, at least in part, be seen through the lens of value chain governance. This casts focus on social dimensions, which have largely been neglected in the certification and standards discussion, such as social conditions throughout the data collection and processing chain. Yet it also casts focus on the limits of certification to change the structural conditions in which machine learning operates, much as sustainability certification cannot replace strong national labor laws or reform the fundamental nature of globalized markets.

Despite this, experiences with sustainability certification have demonstrated that these programs have value in indirect ways. For example, there have been benefits from the development of venues that have promoted the sharing of best practices and have in turn developed useful Codes of Practice, which serve as meta-standards of a sort. Machine learning certification efforts should make an effort to understand the value of spillovers as ways to promote and disseminate best practices outside of the certification program, and as part of a long-term strategy of change. They should also be wary that not all interaction and spillover effects will have a cumulatively "good" impact on the policy challenges. Some may favor incumbents, or introduce steep costs or undesirable socio-technical lock-in.

Hopes that this form of private governance may be able to fully replace other forms of governance are likely misplaced. In the case of machine learning, certification may prove to be most valuable as forum for innovation and learning, proof of concept, the development of knowledge and expertise, and positive interaction between stakeholders, as a stepping-stone to other forms of regulation. For machine learning, given the extraterritorial nature of the commodity involved, the enforcement issues will be even more challenging than they have been in more spatially grounded ones, like agricultural products and textiles. Certainly, data protection authorities are already struggling to provide the expertise and resource needed for oversight of fairness, transparency, and legality of data processing.

As this paper was in the final stages of publication, the European Commission (in April 2021) published a draft Regulation on Artificial Intelligence (the "Artificial Intelligence Act"). Among other provisions, it proposes a certification scheme inspired by and connected to conformity assessment regimes for (primarily) product safety. We do not analyze it in detail here, suffice to say that aspects of it *do* seek to regulate certain credence properties of machine learning systems. Other aspects, such as the enforcement regime, we suspect can strongly benefit from the lessons drawn from the sustainability certification of commodities we explored above. We await what is undoubtedly going to be a long and controversial legislative process with great interest.

Certification as a tool can be an interim step to other forms of regulation and governance. It can provide a niche for technological and policy innovation, a governance system that is more flexible and adaptable than state-based regulation, and which can bring together key stakeholders, as a stepping-stone to other, more universally applicable approaches. Decision-makers and would-be policy entrepreneurs should see certification through this lens: useful toward a wide variety of ends, but no panacea.

Acknowledgments

This paper benefitted from feedback at the PIIRS/Stockholm Resilience Centre workshop at Princeton University, the ECPR Standing Group on Regulatory Governance in Lausanne, from the audiences at the Seoul Policy Tools Workshop at Yonsei University and the AP-PPN Conference at HKUST, as well as reviewers generous with their time during a pandemic.

Data availability statement

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Endnotes

[1] A large number of documents exist to explain machine learning to those from other disciplines and domains, and we point the reader to those if they have not come across the technology before (see e.g. Edwards & Veale 2017; The Royal Society 2017).

[2] A counter example here is the video or board game playing systems developed by companies such as Google DeepMind. These effectively use a simulation of a world (which video games and board games are apt to provide) and use not only human-plays of these systems (such as Go games by champion players), but also the existence of this simulation to play thousands or millions of games automatically. Worlds we cannot accurately model (such as most complex social phenomena) are much less amenable to this type of machine learning training technique.

[3] This is not to be confused with privacy impacts resulting from how that information is used, who has access, or who is disproportionately surveilled.

[4] https://deepempathy.mit.edu/

[5] http://moralmachine.mit.edu/. See further Awad *et al.* (2018).

[6] The only finalized standard at the time of writing is IEEE 7010 *Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being* (available from URL: https://doi.org/10.1109/IEEESTD.2020.9084219).

[7] https://www.isealalliance.org/about-iseal/who-we-are

References

Abbott KW, Levi-Faur D, Snidal D (2017) Introducing Regulatory Intermediaries. *The Annals of the American Academy of Political and Social Science* 670, 6–13.

Agre PE (1994) Surveillance and Capture: Two Models of Privacy. *The Information Society* 10(2), 101–127.

Auld G, Gulbrandsen L (2010) Transparency in Nonstate Certification: Consequences for Accountability and Legitimacy. *Global Environmental Politics* 10, 97–119 https://doi.org/Article.

Auld G, Gulbrandsen LH, McDermott CL (2008) Certification Schemes and the Impacts on Forests and Forestry. *Annual Review of Environment and Resources* 33, 187.

Auld G, Renckens S (2017) Rule-Making Feedbacks through Intermediation and Evaluation in Transnational Private Governance. *The Annals of the American Academy of Political and Social Science* 670, 93–111.

Awad E, Dsouza S, Kim R *et al.* (2018) The Moral Machine Experiment. *Nature* 563(7729), 59.

Barocas S, Selbst AD (2016) Big Data's Disparate Impact. *California Law Review* 104, 671.

Barry M, Cashore B, Clay J *et al.* (2012) *Toward Sustainability: The Roles and Limitations of Certification*, *Final Report*. Steering Committe of the State-of-Knowledge Assessments of Standards and Certification, Washington, DC.

Bender EM, Gebru T, McMillan-Major, A, Shmitchell, S (2021) On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability and Transparency in Computing Systems (ACM FAccT 2021).*

Bennett CJ, Raab CD (2018) Revisiting the Governance of Privacy: Contemporary Policy Instruments in Global Perspective. *Regulation & Governance*.

Bennett EA (2017) Who Governs Socially-Oriented Voluntary Sustainability Standards? Not the Producers of Certified Products. *World Development* 91, 53–69.

Bernstein S, Cashore B (2012) Complex Global Governance and Domestic Policies: Four Pathways of Influence. *International Affairs* 88, 585–604.

Bietti E (2020) From Ethics Washing to Ethics Bashing: A View on Tech Ethics from within Moral Philosophy. *Proceedings of the ACM Conference on Fairness, Accountability and Transparency (ACM FAT* 2020*, Barcelona, Jan 27-30 2020).

Binns R (2018) Fairness in Machine Learning: Lessons from Political Philosophy. *Conference on Fairness, Accountability and Transparency (FAT* 2018), 81*, pp. 1–11. PMLR, New York.

Binns R, Lyngs U, Van Kleek M, Zhao J, Libert T, Shadbolt N (2018) Third Party Tracking in the Mobile Ecosystem. *Proceedings of the 10th ACM Conference on Web Science, WebSci '18*, pp. 23–31. ACM, New York. http://doi.acm.org/10.1145/3201064.3201089.

Bird S, Barocas S, Crawford K, Diaz F, Wallach H (2016) Exploring or Exploiting? Social and Ethical Implications of Autonomous Experimentation in AI, *Paper presented at FAT/ML* 2016. [Last accessed 21 Feb 2019.] Available from URL: https://papers.ssrn.com/abstract=2846909.

Borck JC, Coglianese C (2009) Voluntary Environmental Programs: Assessing their Effectiveness. *Annual Review of Environment and Resources* 34, 305–324.

Brès L, Mena S, Salles-Djelic M-L (2019) Exploring the formal and informal roles of regulatory intermediaries in transnational multistakeholder regulation. *Regulation & Governance* 13(2), 127–140.

Bryson J, Winfield A (2017) Standardizing Ethical Design for Artificial Intelligence and Autonomous Systems. *Computer* 50 (5), 116–119 https://doi.org/10/gfv83v.

Bygrave LA (2001) Minding the Machine: Article 15 of the EC Data Protection Directive and Automated Profiling. *Computer Law and Security Review* 17(1), 17–24 https://doi.org/10/c4j4h2.

Carata L, Akoush S, Balakrishnan N *et al.* (2014) A Primer on Provenance. *Communications of the ACM* 57(5), 52–60.

Cashore B (2002) Legitimacy and the Privatization of Environmental Governance: How Non-state Market-Driven (NSMD) Governance Systems Gain Rule-Making Authority. *Governance-International Journal of Policy and Administration* 15, 503–529.

Cashore B, Auld G, Bernstein S, McDermott C (2007) Can Non-state Governance 'Ratchet Up' Global Environmental Standards? Lessons from the Forest Sector. *Review of European Community and International Environmental Law* 16, 158–172.

Cashore BW, Auld G, Newsom D (2004) *Governing through Markets: Forest Certification and the Emergence of Non-State Authority*. Yale University Press, New Haven, CT.

Cath C, Floridi L (2017) The Design of the Internet's Architecture by the Internet Engineering Task Force (IETF) and Human Rights. *Science and Engineering Ethics* 23, 449–468 https://doi.org/10/f94ggh.

Charlesworth A (2000) Data Privacy in Cyberspace: Not National Vs. International but Commercial Vs. Individual. In: Edwards L, Waelde C (eds) *Law and the Internet: A Framework for Electronic Commerce*. Hart, Oxford.

Chen A (2014) The Laborers Who Keep Dick Pics and Beheadings out of your Facebook Feed. *Wired*. [Last accessed 28 Feb 2019.] Available from URL: https://www.wired.com/2014/10/content-moderation/.

Cohen JE (2019) *Between Truth and Power: The Legal Constructions of Informational Capitalism*. Oxford University Press, Oxford.

Committee of Experts on Internet Intermediaries (MSI-NET) (2017) *Study on the Human Rights Dimensions of Automated Data Processing Techniques (in Particular Algorithms) and Possible Regulatory Implications (MSI-NET(2016)06 Rev3 FINAL)*. Council of Europe, Strasbourg.

Craglia M, Annoni A, Benczur P, Bertoldi P, Delipetrev P, De Prato G *et al.* 2018. *Artificial Intelligence – a European Perspective*. European Commission.

Crawford K, Joler V (2018) Anatomy of an AI System. *anatomyof.ai*. [Last accessed 4 Mar 2019.] Available from URL: https://anatomyof.ai.

Custers BHM, Calders T, Schermer B, Zarsky T (eds) (2013) *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*. Springer, New York.

Cutler AC, Haufler V, Porter T (eds) (1999) *Private Authority and International Affairs*. SUNY Press, Albany, NY.

Darby MR, Karni E (1973) 'Free Competition and the Optimal Amount of Fraud', *The Journal of Law & Economics* 16(1), 67–88.

Edwards L, Veale M (2017) Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For. *Duke Law and Technology Review* 16(1), 18–84.

Ensign D, Friedler SA, Neville S, Scheidegger C, Venkatasubramanian S (2018) Runaway Feedback Loops in Predictive Policing. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (FAT* 2018)*, PMLR 81:160–171.

Ensmenger N (2018) The Environmental History of Computing. *Technology and Culture* October, 59(4):S7–S33.

Gangadharan SP, Niklas J (2019) Decentering Technology in Discourse on Discrimination. *Information, Communication and Society* 22(7), 882–899.

Gellman R, Dixon P (2011) *Many Failures: A Brief History of Privacy Self-Regulation in the United States*. World Privacy Forum, Lake Oswego, OR.

Gereffi G, Humphrey J, Sturgeon T (2005) The Governance of Global Value Chains. *Review of International Political Economy* 12(1), 78–104.

Glickman SW, Cairns CB, Schulman KA (2009) Ethical and Scientific Implications of the Globalization of Clinical Research. *The New England Journal of Medicine* 360(8), 816–823.

Gulbrandsen LH (2004) Overlapping Public and Private Governance: Can Forest Certification Fill the Gaps in the Global Forest Regime? *Global Environmental Politics* 4, 75–99.

Gunningham N (1995) Environment, Self-Regulation, and the Chemical Industry: Assessing Responsible Care. *Law and Policy* 17, 57.

Gunningham N (1996) From Compliance to Best Practice in OHS: The Roles of Specification, Performance and Systems-Based Standards. *Australian Journal of Labor Law* 9(3), 221–243.

Gunningham N, Kagan RA, Thornton D (2004) Social License and Environmental Protection: Why Businesses Go beyond Compliance. *Law & Social Inquiry* 29, 307–341.

Gürses S, van Hoboken J (2018) Privacy after the Agile Turn. In: Selinger E, Polonetsky J, Tene O (eds) *The Cambridge Handbook of Consumer Privacy*, pp. 579–601. Cambridge University Press, Cambridge.

Harcourt A, Christou G, Simpson S (2020) *Global Standard Setting in Internet Governance*. Oxford University Press, Oxford.

Henderson T (2017) *Does the GDPR Help or Hinder Fair Algorithmic Decision-Making?*. University of Edinburgh, LLM.

Hildebrandt M, Gutwirth S (eds) (2008) *Profiling the European Citizen: Cross-Disciplinary Perspectives*. Springer, New York.

Hood C, Rothstein H, Baldwin R (2001) *The Government of Risk: Understanding Risk Regulation Regimes*. Oxford University Press, Oxford.

House of Commons (2018) Algorithms in Decision-Making. Science and Technology Committee (Commons). https://perma.cc/PH52-NUWC.

House of Lords (2018) AI in the UK: Ready, Willing and Able? Select Committee on Artificial Intelligence (Lords).

Huang L, Joseph AD, Nelson B, BIP R, Tygar JD (2011) Adversarial Machine Learning. *Proceedings of AISec '11*. ACM, New York.

Humbert S, Loerincik Y, Rossi V, Margni M, Jolliet O (2009) Life Cycle Assessment of Spray Dried Soluble Coffee and Comparison with Alternatives (Drip Filter and Capsule Espresso). *Journal of Cleaner Production* 17(15), 1351–1358.

Jaffee D (2014) *Brewing Justice: Fair Trade Coffee, Sustainability, and Survival*. University of California Press, Oakland, CA.

Kamara I, de Hert P (2018) Data Protection Certification in the EU: Possibilities, Actors and Building Blocks in a Reformed Landscape. In: Rodrigues R, Papakonstantinou V (eds) *Privacy and Data Protection Seals*. Hague: TMC Asser Press.

Karaj A, Macbeth S, Berson R, Pujol JM (2018) WhoTracks.Me: Shedding light on the opaque world of online tracking. *arXiv*: 1804.08959, accessed May 14, 2019, from http://arxiv.org/abs/1804.08959v2

Keyes O, Hutson J, Durbin M (2019) A Mulching Proposal: Analysing and Improving an Algorithmic System for Turning the Elderly into High-Nutrient Slurry, in: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19) ACM, New York 2019.

King AA, Lenox MJ (2000) Industry Self-Regulation without Sanctions: The Chemical Industry's Responsible Care Program. *Academy of Management Journal* 43(4), 698–716.

King AA, Lenox MJ, Terlaak A (2005) The Strategic Use of Decentralized Institutions: Exploring Certification With the ISO 14001 Management Standard. *Academy of Management Journal* 48(6), 1091–1106.

Komives K, Jackson A (2014) Introduction to Voluntary Sustainability Standard Systems. In: Schmitz-Hoffmann C, Schmidt M, Hansmann B, Palekhov D (eds) *Voluntary Standard Systems: A Contribution to Sustainable Development*, pp. 3–19. Springer, Berlin, Heidelberg.

Kosta E (Forthcoming) *Algorithmic State Surveillance: Challenging the Notion of Agency in Human Rights. Regulation and Governance.*

Kourula A, Paukku M, Peterman A, Koria M (2019) Intermediary roles in regulatory programs: Toward a role-based framework. *Regulation & Governance* 13(2), 141–156.

Kroll J, Huey J, Barocas S *et al*. (2017) Accountable Algorithms. *University of Pennsylvania Law Review* 165(3), 633.

Lambin EF, Thorlakson T (2018) Sustainability Standards: Interactions between Private Actors, Civil Society, and Governments. *Annual Review of Environment and Resources* 43, 369–393.

Leenes R (2020) Article 42. Certification. In: Kuner C, Bygrave LA, Docksey C (eds) *The EU General Data Protection Regulation (GDPR): A Commentary*, pp. 732–743. Oxford University Press, Oxford.

Lenox M (2006) The Role of Private Decentralized Institutions in Sustaining Industry Self-Regulation. *Organization Science* 17 (6), 677.

Lenox M (2007) Do Voluntary Standards Work among Corporations? The Experience of the Chemicals Industry. *Making Global Self-Regulation Effective in Developing Countries*: 62.

Lodge M, Stirton L (2010) Accountability in the Regulatory State. In: Baldwin R, Cave M, Lodge M (eds) *The Oxford Handbook of Regulation*. Oxford University Press, Oxford.

Lum K, Isaac W (2016) To Predict and Serve. *Significance* 13(4), 14–19.

Marsden CT (2011) *Internet Co-Regulation: European Law, Regulatory Governance and Legitimacy in Cyberspace*. Cambridge University Press, Cambridge.

Mattli W, Büthe T (2003) Setting International Standards: Technological Rationality or Primacy of Power? *World Politics* 56 (1), 1–42.

Matus KJM (2010a) Policy Incentives for a Cleaner Supply Chain: The Case of Green Chemistry. *Journal of International Affairs* 64(1), 121–136.

Matus KJ (2010b) Standardization, certification, and labeling: A background paper for the roundtable on sustainability workshop. *Committee on Certification of Sustainable Products and Services (ed), Certifiably Sustainable?* National Academies Press.

Matus KJM (2014) Capacity, Innovation, and their Interaction in Multi-Stakeholder Sustainability Initiatives. In: Lodge M, Wegrich K (eds) *The Problem-Solving Capacity of the Modern State: Governance Challenges and Administrative Capacities*, p. 258. Oxford University Press, Oxford.

McCluskey JJ, Loureiro ML (2005) Reputation and Production Standards. *Journal of Agricultural and Resource Economics* 30 (1), 1–11.

Mikolov T, Chen, K, Corrado, G and Dean, J (2013) Efficient Estimation of Word Representations in Vector Space. *arXiv*: 1301.3781, accessed February 28, 2019, from http://arxiv.org/abs/1301.3781

Miorandi D, Rizzardi A, Sicari S, Coen-Porisini A (2020) Sticky Policies: A Survey. *IEEE Transactions on Knowledge and Data Engineering* 32(12), 2481–2499.

Mitchell TM (1997) *Machine Learning*. McGraw Hill, Burr Hill, IL.

Mol APJ, Oosterveer P (2015) Certification of Markets, Markets of Certificates: Tracing Sustainability in Global Agro-Food Value Chains. *Sustainability* 7, 12258–12278.

Morozov E (2019) Capitalism's New Clothes. *The Baffler*. [Last accessed 28 Feb 2019.] Available from URL: https://thebaffler.com/latest/capitalisms-new-clothes-morozov.

Ingraham N (2013) Larry Page Wants to 'Set Aside a Part of the World' for Unregulated Experimentation. *The Verge*. [Last accessed 2 Mar 2019.] Available from URL: https://www.theverge.com/2013/5/15/4334356/larry-page-wants-to-set-aside-a-part-of-the-world-for-experimentation.

Neilson J, Prichard B (2009) *Value Chain Struggles: Institutions and Governance in the Plantation Districts of South India*. Wiley-Blackwell, London.

Newton C (2019) The Trauma Floor: The Secret Lives of Facebook Moderators in America. *The Verge*. [Last accessed 28 Feb 2019.] Available from URL: https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona.

Parker C, Carey R, Costa JD, Scrinis G (2017) Can the Hidden Hand of the Market Be an Effective and Legitimate Regulator? The Case of Animal Welfare under a Labeling for Consumer Choice Policy Approach. *Regulation and Governance* 11(4), 368–387.

Pennington J, Socher R, Manning C (2014) GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Stroudsburg, PA, pp. 1532–1543.

Petryna A (2009) *When experiments travel: Clinical trials and the global search for human subjects*. Princeton University Press.

Petty T, Saba M, Lewis T, Peña Gangadharan S, and Eubanks V (2018) *Our Data Bodies: Reclaiming Our Data*. [Last accessed 23 Oct 2020.] Available from URL: https://perma.cc/8GF2-GL6X.

Picard RW (1997) *Affective Computing*. MIT Press, Cambridge, MA.

Reidenberg JR (1998) Lex Informatica: The Formulation of Information Policy Rules through Technology. *Texas Law Review* 76, 553–594.

Roberts ST (2019) *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press, New Haven.

Ryan J (2011) *A History of the Internet and the Digital Future*. Reaktion Books, London.

Sassen S (1996) *Losing Control? Sovereignty in the Age of Globalization*. Columbia University Press, New York.

Sauter M (2018) Google's Guinea-Pig City. *The Atlantic*. [Last accessed 2 Mar 2019.] Available from URL: https://www.theatlantic.com/technology/archive/2018/02/googles-guinea-pig-city/552932/.

Seering J, Fang T, Damasco L, Chen M, Sun L, Kaufman G (2019) Designing User Interface Elements to Improve the Quality and Civility of Discourse in Online Commenting Behaviors. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, 1–14. https://doi.org/10.1145/3290605.3300836

Silberman M, Six LI, Ross J (2010) Ethics and Tactics of Professional Crowdwork. *XRDS* 17(2), 39–43.

Singh J, Cobbe J, Norval C (2019) Decision Provenance: Harnessing Data Flow for Accountable Systems. *IEEE Access* 7, 6562–6574.

Taylor L, Dencik L (2020) Constructing Commercial Data Ethics. *Technology and Regulation* 2020, 1–10.

The Royal Society (2017) *Machine Learning: The Power and Promise of Computers that Learn by Example*. The Royal Society, London. https://royalsociety.org/machine-learning.

van der Heijden J (2012) Voluntary Environmental Governance Arrangements. *Environmental Politics* 21, 486–509.

Veale M, Binns R (2017) Fairer Machine Learning in the Real World: Mitigating Discrimination without Collecting Sensitive Data. *Big Data and Society* 4(2), 205395171774353 https://doi.org/10/gdcfnz.

Veale M, Binns R, Edwards L (2018b) Algorithms that Remember: Model Inversion Attacks and Data Protection Law. *Philosophical Transactions of the Royal Society A* 376, 20180083.

Veale M, Seixas R (2015) Moving to Metrics: Opportunities and Challenges of Performance-Based Sustainability Standards. *S. A.P.I.EN.S* 8(1), 1713.

Veale M, Van Kleek M, Binns R (2018a) Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2018)*. ACM, New York https://doi.org/ct4s.

Weigmann K (2015) The ethics of global clinical trials. *EMBO Reports* 16(5), 566–570.

Yates J, Murphy C (2019) *Engineering Rules: Global Standard Setting since 1880*. Johns Hopkins University Press, Baltimore.

Yeung K and Bygrave K (Forthcoming) A Critical Examination of the Legitimacy of the Modernised European Data Protection Regime through a 'Decentred' Regulatory Lens. *Regulation and Governance*.

Zuboff S (2015) Big Other: Surveillance Capitalism and the Prospects of an Information Civilization. *Journal of Information Technology* 30(1), 75–89.

Cases cited

Case C-311/18 *Data Protection Commissioner v. Facebook Ireland and Schrems* ECLI:EU:C:2020:559.

Case C-362/14 *Maximillian Schrems v. Data Protection Commissioner* ECLI:EU:C:2015:650.

Laws cited

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) OJ L 119/1.