Some HCI Priorities for GDPR-Compliant Machine Learning

Michael Veale
University College London
London, United Kingdom
m.veale@ucl.ac.uk

Reuben Binns
Max Van Kleek
University of Oxford
Oxford, United Kingdom
reuben.binns@cs.ox.ac.uk
emax@cs.ox.ac.uk

Abstract

In this short paper, we consider the roles of HCI in enabling the better governance of consequential machine learning systems by the rights and obligations laid out in the recent 2016 EU General Data Protection Regulation (GDPR) which involve heavy interaction with people and systems. Focussing on those areas that relate to algorithmic systems in society, we propose roles for HCI in legal context in relation to fairness, bias and discrimination; data protection by design; data protection impact assessments; transparency and explanations; the mitigation and understanding of automation bias; and the communication of envisaged consequences of processing.

Introduction

The 2016 EU General Data Protection Regulation (GDPR) is making waves. With all personal data relating to EU residents or processed by EU companies within scope, it seeks to strengthen data subject rights and data controller obligations in an increasingly data-laden society, newly underpinned with both an overarching obligation of data controller accountability and hefty maximum fines. Its articles introduce new provisions, and formalise existing rights clarified by the European Court of Justice (such as the “right to be forgotten”) as well as strengthening those already present in the 1995 Data Protection Directive (DPD).

Data Subjects & Controllers
EU DP law applies whenever personal data is processed either in the Union, or outside the Union relating to an EU resident. Personal data is defined by how much it can render somebody identifiable—going beyond email, phone number, etc to include dynamic IP addresses, browser fingerprints or smart meter readings. The individual data relates to is called the *data subject*. The organisation(s) who determine 'the purposes and means of the processing of personal data' are *data controllers*. Data subjects have rights over personal data, such as rights of access, erasure, objection to processing, and portability of data elsewhere. Data controllers are subject to a range of obligations, such as ensuring confidentiality, notifying if data is breached, and undertaking risk assessments. Additionally, they must only process data where they have a legal ground—such as consent—to do so, for a specified and limited purpose, and a limited period of storage.

The GDPR has been turned to by scholars and activists as a tool for "algorithmic accountability" in a society where machine learning (ML) seems to be increasingly important. Machine learning models—statistical systems which use data to improve their performance on particular tasks—are the approach of choice to generate value from the 'data exhaust' of digitised human activities. Critics, however, have framed ML as powerful, opaque, and with potential to endanger privacy [2], equality [10] and autonomy [20]. While the GDPR is intended to govern personal data rather than ML, there are a range of included rights and obligations which might be useful to exert control over algorithmic systems [14].

Given that GDPR rights involve both individual data-subjects and data controllers (see sidebar) interfacing with computers in a wide variety of contexts, it strongly implicates another abbreviation readers will likely find familiar: Human–Computer Interaction (HCI). In this short paper, we outline, non-exhaustively of course, some of the cross-overs between the GDPR provisions, HCI and ML that appear most salient and pressing given current legal, social, and technical debates. We group these in two broad categories: those which primarily concern the building and training of models *before deployment*, and those which primarily concern the *post-deployment application* of models to data subjects in particular situations.

HCI, GDPR and Model Training

An increasing proportion of collected personal data[1] is used to train machine learning systems, which are in turn used to make or support decisions in a variety of fields. As model training with personal data is considered data processing (assuming data is not solidly 'anonymised'), the GDPR does govern it to a varying degree. In this section, we consider to what extent HCI might play a role in promoting the governance of model training under the GDPR.

Fairness, discrimination and 'special category' data
Interest in unfair and/or illegal data-driven discrimination has concerned researchers, journalists, pundits and policy-makers [17, 3], particularly as the ease of transforming seemingly non-sensitive data into potentially damaging, private insights has become apparent [9]. Meanwhile, most focus on how to govern data (both in Europe and elsewhere broadly [19]) has been centred on *data protection*, in particular, the GDPR. The GDPR does not have anti-discrimination as a core concept—it is not an anti-discrimination law—yet it does have provisions which concern particularly sensitive attributes of data.

Several types of data are "special" in the GDPR, and require higher protection. The 1995 Data Protection Directive (art 8) prohibits processing of data *revealing* **racial or ethnic origin**, **political opinions**, **religious or philosophical beliefs** and **trade-union membership**, in addition to data concerning **health** or **sex life**. The GDPR (art 9(1)) adds **genetic** and **biometric** data (the latter for the purposes of identification), as well as clarifying sex life includes orientation, to create 8 'special categories' of data. This list is similar, but not identical, to the 'protected characteristics' in many international anti-discrimination laws. Compared to the UK's Equality Act 2010, the GDPR omits age, sex and marital status but includes political opinions, trade union membership, and health data more broadly.

The collection, inference and processing of special category data triggers both specific provisions (e.g. arts 9, 22) and specific responsibilities (e.g. Data Protection Impact Assessments, art 35 and below), as well as generally

[1] Note that the GDPR defines personal data broadly—including things like dynamic IP addresses and home energy data—as opposed to the predominantly American notion of personally identifiable information (PII) [25].

heightening the level of risk of processing and therefore the general responsibilities of a controller (art 24). Perhaps the most important difference is that data controllers cannot rely on their own *legitimate interests* to justify the processing of special category data, which usually will mean they will have to seek explicit, specified consent for the type of processing they intend—which they may not have done for their original data, and may not be built into their legal data collection model.

Given that inferred special category data is also characterised as special category data [28], there are important questions around how both controllers and regulators recognise that such inference is or might be happening. Naturally, if a data controller trains a supervised model for the purpose of inferring a special category of data, this is quite a simple task (as long as they are honest about it). Yet when they are using latent characteristics, such as through principal components analysis, or features that are embedded within a machine learning model, this becomes a more challenging task. In particular, it is very possible (and repeatedly shown) how biases based on special category data can appear in trained systems even if those special categories are not present in the datasets being used [9].

The difficulty of this task is heightened by how the controller in this case is unlikely to possess 'ground truth' data in order to assess what it is they are picking up. HCI might play an important role here in establishing what has been described as 'exploratory fairness analysis' [27]. The task is to understand potential patterns of discrimination, or to identify certain unexpected but sensitive clusters, with only partial additional information abut the participants. A similar proposal (and prototype of) a visual system, albeit one assuming full information, has been proposed by discrimination-aware data mining researchers concerned that the formal statistical criteria for non-discrimination established by researchers may not connect with ideas of fairness in practice [4, 5]. If we do indeed also know unfairness when we see it, exploratory visual analysis may be a useful tool. A linked set of discussions have been occurring in the information visualisation community around desirable characteristics of feminist data visualisation, which connects feminist principles around marginalisation and dominance in the production of knowledge to information design [11]. Finally, visual tools which help identify misleading patterns in data, such as instances of Simpson's paradox (e.g. [24]), may prove useful in confirming apparent disparities between groups. Building and testing interfaces which help identify sensitive potential correlations and ask critical questions around bias and discrimination in the data is an important prerequisite to rigorously meeting requirements in the GDPR.

Upstream provisions: Data Protection by Design (DPbD) and Data Protection Impact Assessments (DPIAs)
The GDPR contains several provisions intended to move considerations of risks to data subjects' rights and freedoms upstream into the hands of designers. Data Protection by Design (DPbD), a close cousin of privacy by design, is a requirement under the GDPR and means that controllers should use organisational and technical measures to imbue their products and processes with data protection principles [8]. Data Protection Impact Assessments (DPIAs) have a similar motivation [6]. Whenever controllers have reason to believe that a processing activity brings high risks, they must undertake continuous, documented analysis of these, as well as any measures they are taking to mitigate.

The holistic nature of both DPbD and DPIAs is emphasised in both the legal text and recent guidance. These are creative processes mixing anticipation and foresight with

best practice and documentation. Some HCI research has already addressed this in particular. Luger et al. [23] use *ideation cards* to engage designers with regulation and co-produce "data protection heuristics".[2] Whether DPIA aides can be built into existing systems and software in a user-centric way is an important area for future exploration.

Furthermore, many risks within data, such as bias, poor representation, or the picking up of private features, may be unknown to data controllers. Identifying these is the point of a DPIA, but subtle issues are unlikely to leap out of the page. In times like this, it has been suggested that a shared knoweldgebase [27] could be a useful resource, where researchers and data controllers (or their modelling staff) could log risks and issues in certain types of data from their own experiences, creating a resource which might serve useful in new situations. For example, such a system might log found biases in public datasets (particularly when linked to external data) or in whole genres of data, such as Wi-Fi analytics or transport data. Such a system might be a useful starting point for considering issues that may otherwise go undetected, and for supporting low-capacity organisations in their responsible use of analytics. From an HCI perspective though, the design of such a system presents significant challenges. How can often nuanced biases be recorded and communicated both clearly and in such a way that they generalise across applications? How might individuals easily search a system for issues in their own datasets, particularly when they might have a very large number of variables in a combination the system has not seen previously? Making this kind of knowledge accessible to practitioners seems promising, but daunting.

[2]The cards are downloadable at https://perma.cc/3VBQ-VVPQ.

HCI, GDPR and Model Application

The GDPR, and data protection law in general, was not intended to significantly govern decision-making. Already a strange law in the sense that it poses transparency requirement that applies to the public and private sectors alike, it is also a Frankenstein's monster–style result culminating from the melding of various European law and global principles that preceded it [16].

Modes of Transparency

While transparency is generally spoken of as a virtue, the causal link between it and better governance is rarely simple or clear. A great deal of focus has been placed on the so-called "right to an explanation", where a short paper at a machine learning conference workshop [18] gained sudden notoriety, triggering reactions from lawyers and technologists noting that the existence and applicability of such a right was far from simple [29, 14]. Yet the individualised transparency paradigm has rarely provided much practical use for data subjects in their day-to-day lives (consider the burden of 'transparent' privacy policies). Consequently, HCI provides a useful place to start when considering how to make the limited GDPR algorithmic transparency provisions useful governance tools.

There are different places in which algorithmic transparency rights can be found in the GDPR [29]. Each bring different important HCI challenges.

Meaningful information about the logic of processing

Under the most widely accepted scholarly interpretation, articles 13–14 oblige data controllers to provide information at the time data is collected around the logics of certain automated decision systems that might be applied to this data. Current regulatory guidance [1] states that there is no obligation to tailor this information to the specific situation of

a data subject (other than if they might be part of a vulnerable group, like children, which might need further support to make the information meaningful). This points to an important HCI challenge in making (or visualising) such general information, but with the potential for specific relevance to individuals.

Right to be informed In addition, there is a so-called 'right to be informed' of automated decision-making [29]: how might an interface seamlessly flag to users when a potentially legally relevant automated decision is being made? This is made more challenging by the potential for adaptive interfaces or targeted advertising to meet the criteria of a 'decision'. In these cases, it is unclear at what point the 'decision' is being made, even though the system as a whole might be challenged on the basis it is decision-making with regards to selective information provision or omission. Exercise of data protection rights is different in further ways in ambient environments [13], as smart cities and ambient computing may bring significant challenges, if, for example, they are construed as part of decision-making environments. Existing work in HCI has focussed on the difficulties in identifying "moments of consent" in ubiquitous computing [22, 21]. Not only is this relevant when consent is the legal basis for an automated decision, but additional consideration will be needed in relation to what equivalent "moments" of objection might look like. Given that moments to object likely outnumber moments to consent, this might pose challenges.

A right to an explanation? The popular "right to an explanation" of specific decisions after they have happened sits in a non-binding recital in the GDPR [29], and thus its applicability and enforceability in practice is questionable. However, there is support for a parallel right in varying forms in certain other laws, such as French administrative law or the Council of Europe Convention 108 [15], and HCI researchers have already been testing different explanation facilities proposed by machine learning researchers in qualitative and quantitative settings to see how they compare in relation to different notions of procedural justice [7]. Further research on explanation facilities in-the-wild would be strongly welcome, given that most explanation facilities to date have focussed on the user of a decision-support system rather than an individual subject to an automated decision.

Mitigating Automation Bias

A key trigger condition for the automated decision-making provisions in the GDPR (art 22) [14] centres on the degree of automation of the process. Significant decisions "based solely on automated processing" require at least consent, a contract or a basis in member state law. Recent regulatory guidance indicates that there must be "meaningful" human input undertaken by somebody with "authority and competence" who does not simply "routinely apply" the outputs of the model in order to be able to avoid contestation or challenge [28]. Automation bias has long been of interest to scholars of human factors in computing [26, 12] and the GDPR provides two core questions for HCI in this vein.

Firstly, this setup implies that systems that are expected to outperform humans must always be considered "solely" automated. If a decision-making system is expected to legitimately outperform humans, then meaningful input seems very difficult, any routine disagreement would be at best arbitrary and at worst, harmful. In this case, this serves as yet another (legal) motivating factor to create systems where human users can augment machine results. Even if this proves difficult, when users contest an automated decision under the GDPR, they have a right to human

review. Interfaces need to ensure that even where models may be complex and high-dimensional, decision review systems are rigourous and themselves have “meaningful” human input—or else these reviewed decisions are equally open to contestation.

Secondly, how might a data controller or a regulator understand whether systems have “meaningful” human input or not, in order to either obey or enforce the law? How might this input be justified and documented in a useful and user-friendly way which could potentially be provided to the subject of the decision? Recent French law does oblige this in some cases: in the case of algorithmically-derived administrative decisions, information should be provided to decision-subjects on the “ the degree and the mode of contribution of the algorithmic processing to the decision-making” [15]. Purpose-built interfaces and increased knowledge from user studies both seem needed for the aim of promoting meaningful, accountable input.

Communicating Envisaged Consequences
Where significant, automated decision-making using machine learning is expected, the information rights in the GDPR (arts 13–15) provide that a data subject should be provided with the “envisaged consequences” of such decision for her. What this means is far from clear. Recent regulatory guidance provides only the example of giving data subject applying for insurance premiums an app to demonstrate the consequences of dangerous driving [1]. Where users are consenting to complex online personalisation which could potentially bring significant effects to their life, such as content delivery which might lead to echo chambers or “filter bubbles”, it is unclear how complex “envisaged consequences” might be best displayed in order to promote user autonomy and choice.

Concluding remarks

HCI is well-placed to help enable the regulatory effectiveness of the GDPR in relation to algorithmic fairness and accountability. Here we have touched on different points where governance might come into play—model training and model application—but also different modes of governance. Firstly, HCI might play a role in enabling creative, rigorous, problem solving practices within organisations. Many mechanisms in the GDPR, such as data protection by design and data protection impact assessments, will depend heavily on the communities, practices and technologies that develop around them in different contexts. Secondly, HCI might play a role in enabling controllers do particular tasks better. Here, we discussed the potential for exploratory data analysis tools, such as detecting special category data even when it was not explicitly collected. Finally, it might help data subjects exercise their rights better. It appears especially important to develop new modes and standards for transparency, documentation of human input, and communication of tricky notions such as “envisaged consequences”.

As the GDPR often defines data controllers’ obligations as a function of “available technologies” and “technological developments”, it is explicitly enabled and strengthened by computational systems and practices designed with its varied provisions in mind. Many parts of the HCI community have already been building highly relevant technologies and practices that could be applied in this way. Further developing these with a regulatory focus might be transformative in and of itself—and it something we believe should be promoted in this field and beyond.

REFERENCES

1. Article 29 Data Protection Working Party. 2017. *Draft Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, wp251*.

2. Solon Barocas and Helen Nissenbaum. 2014. Big Data's End Run Around Procedural Privacy Protections. *Commun. ACM* 57, 11 (2014), 31–33.

3. Solon Barocas and Andrew D Selbst. 2016. Big Data's Disparate Impact. *California Law Review* 104 (2016), 671–732.

4. Bettina Berendt and Sören Preibusch. 2012. Exploring discrimination: A user-centric evaluation of discrimination-aware data mining. In *12th IEEE International Conference on Data Mining Workshops (ICDMW)*. 344–351.

5. Bettina Berendt and Sören Preibusch. 2014. Better decision support through exploratory discrimination-aware data mining: Foundations and empirical evidence. *Artificial Intelligence and Law* 22, 2 (2014), 175–209.

6. Reuben Binns. 2017. Data protection impact assessments: A meta-regulatory approach. *International Data Privacy Law* 7, 1 (2017), 22–35. DOI:http://dx.doi.org/10.1093/idpl/ipw027

7. Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage'; Perceptions of Justice in Algorithmic Decisions. *Proceedings of the SIGCHI conference on Human Factors in Computing Systems (CHI'03)* (2018).

8. Lee A Bygrave. 2017. Data Protection by Design and by Default: Deciphering the EU's Legislative Requirements. *Oslo Law Review* 1, 02 (2017), 105–120.

9. Toon Calders and Indrė Žliobaitė. 2012. Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures. In *Discrimination and Privacy in the Information Society*, Bart Custers, Toon Calders, Bart Schermer, and Tal Zarsky (Eds.). Springer, Berlin, Heidelberg, 43–59.

10. Bart Custers. 2012. Data Dilemmas in the Information Society: Introduction and Overview. In *Discrimination and privacy in the information society*, Bart Custers, Toon Calders, Bart Schermer, and Tal Zarsky (Eds.). Springer, 3–25.

11. Catherine D'Ignazio and Lauren F Klein. 2016. Feminist data visualization. In *Workshop on Visualization for the Digital Humanities (VIS4DH), IEEE VIS 2016*.

12. Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. 2003. The role of trust in automation reliance. *International Journal of Human-Computer Studies* 58, 6 (2003), 697–718.

13. Lilian Edwards. 2016. Privacy, security and data protection in smart cities: A critical EU law perspective. *Eur. Data Prot. L. Rev.* 2 (2016), 28.

14. Lilian Edwards and Michael Veale. 2017. Slave to the Algorithm? Why a 'Right to an Explanation' is Probably Not The Remedy You Are Looking For. *Duke Law & Technology Review* 16, 1 (2017), 18–84. DOI: http://dx.doi.org/10.2139/ssrn.2972855

15. Lilian Edwards and Michael Veale. 2018. Enslaving the algorithm: From a 'right to an explanation' to a 'right to better decisions'? *IEEE Security & Privacy* (2018). DOI:http://dx.doi.org/10.2139/ssrn.3052831

16. Gloria González Fuster. 2014. *The emergence of personal data protection as a fundamental right of the EU*. Springer.

17. Oscar H Gandy. 2009. *Coming to Terms with Chance: Engaging Rational Discrimination and Cumulative Disadvantage*. Routledge, London.

18. Bryce Goodman and Seth Flaxman. 2016. European Union regulations on algorithmic decision-making and a "right to explanation". In *2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY*.

19. Graham Greenleaf. 2018. 'European' Data Privacy Standards Implemented in Laws Outside Europe. *Privacy Laws & Business International Report* 149 (2018), 21–23. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3096314

20. Mireille Hildebrandt. 2008. Defining Profiling: A New Type of Knowledge? In *Profiling the European Citizen: Cross-Disciplinary Perspectives*, Mireille Hildebrandt and Serge Gutwirth (Eds.). Springer, 17–45.

21. Ewa Luger and Tom Rodden. 2013a. An Informed View on Consent for UbiComp. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '13)*. 529–538. DOI: http://dx.doi.org/10.1145/2493432.2493446

22. Ewa Luger and Tom Rodden. 2013b. Terms of agreement: Rethinking consent for pervasive computing. *Interacting with Computers* 25, 3 (2013), 229–241.

23. Ewa Luger, Lachlan Urquhart, Tom Rodden, and Michael Golembewski. 2015. Playing the legal card: Using ideation cards to raise data protection issues within the design process. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems (CHI'15)*. 457–466. DOI: http://dx.doi.org/10.1145/2702123.2702142

24. Justin Matejka and George Fitzmaurice. 2017. Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems (CHI'17)*. 1290–1294.

25. Paul M Schwartz and Daniel J Solove. 2011. The PII problem: Privacy and a new concept of personally identifiable information. *New York University Law Review* 86 (2011), 1814.

26. Linda J Skitka, Kathleen L Mosier, and Mark Burdick. 1999. Does automation bias decision-making? *International Journal of Human-Computer Studies* 51 (1999), 991–1006. DOI: http://dx.doi.org/10.1006/ijhc.1999.0252

27. Michael Veale and Reuben Binns. 2017. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society* 4, 2 (2017). DOI: http://dx.doi.org/10.1177/2053951717743530

28. Michael Veale and Lilian Edwards. 2018. Clarity, Surprises, and Further Questions in the Article 29 Working Party Draft Guidance on Automated Decision-Making and Profiling. *Computer Law & Security Review* (2018). DOI: http://dx.doi.org/10.1016/j.clsr.2017.12.002

29. Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. 2017. Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law* 7, 2 (2017), 76–99.