Like circles in the water: Responsibility as a system-level function

Sileno, G.; Boer, A.; Gordon, G.; Rieder, B.

Publication date
2021
Document Version
Final published version
Published in
Proceedings of the 3rd EXplainable AI in Law Workshop (XAILA 2020)

Like Circles in the Water: Responsibility as a System-Level Function

Giovanni Sileno[1], Alexander Boer[2], Geoff Gordon[1], and Bernhard Rieder[1]

[1] University of Amsterdam, Amsterdam, the Netherlands
[2] KPMG, Amsterdam, the Netherlands

Abstract. What eventually determines the semantics of algorithmic decision-making in not the program artefact, nor—if applicable—the data used to create it, but the preparatory (enabling) and consequent (enabled) practices holding in the environment (computational and human) in which such algorithmic procedure is embedded. The notion of responsibility captures a very similar construct: in all human societies actions are evaluated in terms of the consequences they could reasonably cause, and of the reasons that motivate them. But to what extent does this function exist in computational systems? The paper aims to sketch links between several of the approaches and concepts proposed for *responsible computing*, from AI to networking, identifying gaps and possible directions for operationalization.

Keywords: Responsibility · Responsible Computing · Responsible AI; Responsible Networking · Contextual Integrity · Conditional Contextual Disparity.

1 Introduction

The various emerging research tracks denoted as *responsible*, *ethical*, *fair*, and *trustworthy AI* can be overall divided in two main families. On the one hand, works contributing to the discussion of what (ethical) principles should be applied, in all phases from conception to deployment, to algorithmic decision-making systems. On the other, works attempting to operationally define open concepts as e.g. "fairness" or "privacy" to be embedded during training or deployment of AI modules. The distance existing between these two approaches raises critical concerns on whether they can be bridged at all. This paper argues for a change of perspective. What eventually determines the semantics of algorithmic decision-making is not the program artefact in itself, nor the data used to create it, but consists of preparatory (enabling) and consequent (enabled) practices holding in the environment in which the algorithmic procedure is embedded. In parallel work [12], we are exploring methods to investigate how "values" are generated, distributed, and translated between contextualized social processes

*This research was partly supported by NWO (DL4LD project, no. 628.001.001) and the RPA Human(e) AI seed grant funded by the UvA.

and automatic/automated decision-making components; inspired by the idea of *encircling* introduced in security studies [3], we are studying how to approach *de facto* inaccessible or opaque entities by looking at what is occurring in their background (practices, ambient knowledge, etc.). The present paper, instead, is meant to take a position in the debate concerning the *system-design* part of the problem. Even acknowledging the primacy of (highly contextual and dynamic) human factors in setting the premises and the consequences of the system's activity, system designers and developers still need solutions to identify and reduce frictions deemed (or feared) to occur between computational and societal dimensions. With this requirement in mind, the paper organizes insights coming from different domains, aiming to be "minimally complete" in highlighting the functions required to achieve a sound infrastructure for responsible computing.

The paper proceeds as follows. Section 1 contrasts a *data-flow* perspective against the most common data-centric ones. Section 2 reviews under a data-flow perspective two non-technical frameworks highlighting the role of context: *contextual integrity* [10], and *contextual demographic disparity* [16]. Section 3 shortly elaborates on the function and functioning of *responsibility* as a cognitive mechanism. Section 4 considers a recent proposal on *responsible Internet* [8] revisiting the *accountability-responsibility-transparency* (ART) principles for AI [4] in the domain of networking, and elaborates on how extending it to take into account what presented in the previous sections.

2 From data to data-flow problems

Most approaches emerging in responsible AI and related fields with respect to problems of *fairness* (non-discrimination) focus primarily on selecting or producing adequate data. Following the overview given in [6], one can for instance:

1. purge the input data from sensitive elements at runtime,
2. debias the sample data used during the training process,
3. correct the network parameters used in the inferential model, or
4. add an external module to produce unbiased output at aggregate level.

These interventions can be interpreted in terms of *computational reflection*, i.e. the ability of a system to inspect and modify itself in order to improve its performance (see e.g. [1]), generally further distinguished in: (a) *structural reflection*, concerned by non-contingent properties of the system (e.g. data structures, procedures); (b) *behavioural reflection*, concerned by the overall activity of the system, as described e.g. by requests/invocations. Using these definitions, options 1, 2, 4 become examples of behavioural reflection: they introduce additional modules invoked to process the input before and/or the output after the core module, without modifying it structurally; 3 is instead an example of structural reflection (it concerns the neural network parameters). In all cases the focus is on *data* (either input, output or relative to the model): even behavioural reflection does not use any information beyond which types of data are protected.

Alternatively, one can see fairness as a problem of *data-flow*: i.e. of intervening or constraining adequately the connections existing between the data processing components. Some of these connections are deemed to be legitimate, others are not; *when illegitimate, the informational connection needs to be cut*, or, at least, to be intervened upon. This change of perspective facilitates the convergence of various problems into one of *responsible processing of informational streams*. Privacy can be seen a set of limited rights and abilities controlling disclosure-of (i.e. channels transmitting) self-information. *Differential privacy* methods [5], introduced to protect against the reconstruction of data of individuals by intersection of a sufficient number of queries, work by adding external noise channels, destroying part of the information by interference. Furthermore, not all applications of "discrimination" (in the sense of distinguishing, characterizing) are negative; they can also bring a positive impact on the data subjects and on society. Initiatives as those driven by the FAIR principles e.g. in healthcare, implicitly support the construction of informational connections. To summarize, it is not only a matter of responsible machine learning, but of *responsible computing* (including processing, data-sharing, networking, etc.). At functional level, a data-flow perspective highlights the pivotal role of the **control of information disclosure**, which can be ***negative*** (i.e. restricting, limiting disclosure) or ***positive*** (i.e. enabling, granting it).

3 The role of context

At face value, technical solutions as those proposed for *algorithmic fairness* or *differential privacy* tend to focus on internal components or the very first layer beyond the system boundaries (input/output data). However, the legitimacy of a certain query or computation is not a problem of the processing in itself, but of the context in which such a processing is performed. For instance, the use of sensitive data such as ethnicity (or proxies of it) is deemed unfair in tasks that produce effects of social discrimination (e.g. deciding the premium for an insurance policy), but not necessarily in other tasks (e.g. deciding the colour/style of a dress in an e-shop). As a paradoxical situation, would we need differential privacy when we are querying our own personal data? More in detail, interventions for algorithmic fairness are meant primarily for three purposes [2]:

- *anti-classification*: decisions are taken without considering explicitly sensitive or protected attributes (ethnicity, gender, etc. or any proxies of those);
- *classification parity*: performance of prediction as measured e.g. by false positive and false negative rates are equal across the groups selected by protected attributes;
- *calibration*: outcomes of prediction is independent of protected attributes.

These purposes reflect in distinct definitions that are incompatible amongst each other, and, furthermore, they can produce effects which are still detrimental to the protected classes [2]. Then, even at a technical level, it is recognized that something is missing in the picture.

The well-known framework of *contextual integrity* by Nissenbaum [10] makes clear that privacy can not be defined in absolute terms, but depends on several parameters, including the actors involved (data subject, sender, recipient), the type of information, the basis for disclosure/transmission, *and* various contextual elements. For instance, consent acts as a basis for disclosure of personal data (e.g. biometrical information) for a specific purpose (e.g. healthcare research), and any other use (e.g. marketing) would be a breach of contextual integrity. However, in some cases (e.g. for medical necessity), the processing of the same personal data without consent will not count as a breach of contextual integrity, because there are legal or even moral norms making clear the presence of a situation (e.g. where survival is at stake) providing a distinct basis for disclosure. In general, context is not defined only by purpose, but also by *domain knowledge* associated with that purpose in the current situation (e.g. norms and practices, and roles related to those), and that is used by the subject and other parties to form their expectations. It is the ecological nature of all these contextual elements that make difficult if not impossible to captured them *monistically* within the informational artefacts which are target of directives about disclosure.

Recent work by Wachter et al. [16] analyzes the concept of *contextual demographic (dis)parity* (CDD) (based on the measure of *conditional (non-)discrimination* proposed by Kamiran et al. in [9]), evaluating it with respect to the decisions of the European Court of Justice on cases of discrimination. The authors highlight the complexity of automatizing decisions about discrimination and suggest therefore to separate (a) the assessment of automated discrimination (and argue that the best measure for this is CDD) from (b) the actual judicial interpretation. Their argument can be rephrased in behavioural reflection terms: the authors are identifying a larger coverage of the network that can be explored by algorithmic-driven assessment, but still make clear that further layers exist beyond that, and this fact requires to maintain human experts in the decision-making loop.

Let us have a further look at CDD. Suppose a norm aims to protect certain groups of people, and suppose a certain decision process produces a positive or negative outcome, dividing people whose data is under scrutiny in two classes, *advantaged* and *disadvantaged*. The authors propose that a *prima facie* assessment of discrimination can be expressed if $A_R < D_R$ for any R in a given set of conditions, where A_R is the proportion of people with protected attributes in the advantaged class, D_R is the proportion of people with protected attribute in the disadvantaged class, and R are additional conditions used to divide the population into sub-classes. But how to decide R? Following Kamiran [9], these conditions should be *explanatory*, i.e. they should hypothetically explain the outcome even in the absence of discrimination against the protected class. For instance, a reason for different salaries between men and women might be different working hours. Indeed, as argued by Pearl [11], the only way out of Simpson's paradox (opposite conclusions using different granularity of observation) is to deal with *causation*. However, questions about "what caused what" have also a strong connection with the idea of *responsibility*. This suggests that other elements may

be needed to the picture in order to evaluate the "reverberations" of the agents' actions onto the system.

4 Function and types of responsibility

Human communities exhibit ascription of responsibility as a spontaneous, seemingly universal behaviour. On an abstract level, responsibility attribution is functional to the *localization* of *failures* in constructions whose components are deemed to be *autonomous*. This construct applies not only to social systems, but to any type of system (natural, artificial, etc.), as it is prerequisite to properly implement remedy/repair function (cf. the *single-responsibility* design principle: one module encapsulates one functionality). Yet, we need to distinguish at least two dimensions of responsibility: *causal* (physical, technical, operational, ...) responsibility, from *moral* (legal, social, ...) responsibility.

Causal responsibility is meant to identify which ones, amongst the components involved in a chain of events, *actually caused* (or prevented) a certain outcome). It generally builds upon properties as *counterfactuality*, *sufficiency* or concurrency. *Moral responsibility* builds upon causal responsibility (although in some circumstances it over-determines it), but it also presupposes a preferential or value structure about possible outcomes in the world: *blame* or *praise* would not make sense for morally irrelevant outcomes.

Empirical studies (e.g. [13], for a unifying computational model see e.g. [15]) suggest that moral responsibility: (i) may generally hold for actions merely initiating potential causes of an outcome; (ii) grows with the impact of the outcome in terms of a preferential/value structure; (iii) is diminished e.g. if the action is not under the (expected) control of the agent, or the outcome is (justifiably) not foreseeable from the agent standpoint.

Rather than facing the question of what makes an agent a moral agent, we can more conservatively identify three requirements for assessing **agentive responsibility**:

1. the agent has the *ability to control* its behaviour;
2. it has the *ability to foresee* the associated outcomes;
3. it has the *ability to assess* their impact according to a preferential/value structure.

None of these three abilities can be absolute. In general, they can be attributed to any (direct and indirect) participants of an interaction, depending on their characteristics and role in the processing network. Furthermore, they are all context dependent—and the definition of context may not be consistent across observers. Note that foreseeability and assessment of impact play a central role in formulating *risk*.

If responsibility is concerned primarily by actions (or activities), **accountability** is generally seen as concerned by providing reasons and justifying those actions (or their omission). Additionally, the occurrence of unmet shared expectations might entail consequences, especially in the presence of a (semi-

)formalized system of norms: **liability** refers to potential duties (e.g. paying damages) associated to those failures, or to other special contexts.

5 Operationalizing responsible computation

Several contributions in the field of *ethical AI* have presented a number of principles for the design and deployment of artificial devices. Consider for instance the ART principles proposed by Dignum [4]: *accountability*: motivations for the decision-making (values, norms, etc.) need to be explicit; *responsibility*: the chain of (human) control (designer, manufacturer, operator, etc.) needs to be clear; *transparency*: actions need to be explained in terms of algorithms and data, and it should be possible to inspect them. However, there is no framework bridging those higher-level principles to the abstraction level of technical solutions as e.g. algorithmic fairness and differential privacy. Impediments can be identified both on a societal dimension (explicit power allocations are conflictual in nature) and from an operational point of view (e.g. policies are expressed at different levels of abstraction, are dynamic, etc.). Additionally, those higher-level proposals tend to look at technological artefacts as essentially monolithical.

Interestingly, a recent paper by Hesselman et al. on the concept of *responsible Internet* [8] takes an orthogonal view over this matter, both in terms of operationalization, and of decentralization. The authors do not focus on the processing of data for decision-making, but on its transmission across the network (cf. the data-flow view), a task that needs to be solved on a decentralized architecture with distributed ownership and control. The paper revisits and slightly modifies the ART principles [4], inflecting them on the dimensions of data and infrastructure. For instance, *data transparency* holds if the system is able to describe how network operators transport and process a certain data-flow, whereas *infrastructure transparency* concerns instead the properties and relationships between network operators (location, software, servers, etc.). The same distinction applies to accountability. Instead of responsibility, however, Hesselman et al. prefer to refer to *controllability*, to focus more on the ability of users to specify how network operators should handle their data (generally by means of *path control*), and to the ability of infrastructure maintainers to set constraints over network operators.[3]

How this more technical view on responsibility relates with the properties of responsibility sketched in the previous section? Accountability and transparency are instrumental to the ascription of responsibility in the moment of failure; they refer to two distinct standpoints over the investigated component, respectively at *functional/extra-functional* levels (accountability), and *non-functional* or implementation level (transparency). The choice of the concept of "controllability" rather than "responsibility" highlights the requirement of setting up the control structure that enables licit outcomes, and prevents illicit outcomes to occur. As

[3] Additionally, they introduce the *usability* principle: the working of the system needs to be expressed in a way that enables further analysis (a practical requirement impacting both transparency and accountability).

we saw in the previous sections, however, (computational) agentive responsibility is not only a matter of controllability, but also of foreseeability, and of the ability of the agent of assessing foreseen outcomes in terms of a given preferential/value structure. Even if the preferential/value structure (of the user, infrastructure maintainer, etc.) can be considered to be part of the input exploiting controllability, the picture implicitly misses the contextual domain knowledge necessary for the agent to make a proper judgement, and that users will seldom have. To correct this, each agent (e.g. a network operator) should in principle autonomously assess its own and other agents' conduct, informed by (i) user policies and norms, (ii) known and potentially relevant scenarios (together with some information about their relative occurrence), attempting to form a properly grounded *risk assessment*.[4] In this view, solutions for algorithmic fairness or differential privacy would be controlled instrumentally to reduce dynamically identified risks.[5] Interestingly, the "distributed responsibility" sketched here is also hinted to in modern legislation as the GDPR, as for instance in Art. 28, according to which the data processor is not any more a mere executor, but it has responsibility that the processing requested by the data-controller is complying with the rules.

Conclusion

The paper results from an effort to organize insights coming from different disciplines and domains related to the topic of *responsible computing*. The bottom line of our investigation is that, in contrast to the most common view taken today in technical approaches, issues like privacy and fairness refer to context-dependent and plural norms (where norm is used as in normative, and as in normality, cf. the concept of *normware* [14]), that cannot be directly translated to optimization tasks. Not all bias is unfair, it depends on how it is used and for what. Not all disclosure is illicit; in fact, some might be beneficial to the data subject and to society. To protect against misuses and improvident disclosures, and thus to achieve responsible computing, computation needs to be looked at in distributed terms (including the associated human activities), and computational agents need to be furnished with some degree of autonomy to be able to assess independently, on the basis of (plural) directives given by humans

[4] Similar considerations apply looking beyond the technological boundaries, cf. Helberger et al. [7] with the concept of "*cooperative responsibility*". In principle, observability should be spread more widely over e.g. civil society actors and not merely individuals and regulators.

[5] In many aspects the term "risk" has already a prominent role in governance technology. However, as it has been observed by several authors (e.g. Rouvroy, Dillon, etc.) the alignment of risk analysis with competitive value extraction contributes to a very particular policy platform which is not neutral. These critics do not make risk a necessarily illegitimate category, but point to ways to further elaborate the importance of context, including specific contextual features to acknowledge policy concerns going beyond value extraction.

and (plural) knowledge constructed from system practices, whether a certain requested processing is indeed justified.

References

1. Capra, L., Blair, G.S., Mascolo, C., Emmerich, W., Grace, P.: Exploiting reflection in mobile computing middleware. ACM SIGMOBILE Mobile Computing and Communications Rev. **6**(4), 34–44 (10 2002)
2. Corbett-Davies, S., Goel, S.: The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning (2018), http://arxiv.org/abs/1808.00023
3. De Goede, M., Bosma, E., Pallister-Wilkins, P.: Secrecy and Methods in Security Research: A Guide to Qualitative Fieldwork. Routledge (2019)
4. Dignum, V.: Responsible autonomy. Proceedings of International Joint Conference on Artificial Intelligence (IJCAI) pp. 4698–4704 (2017)
5. Dwork, C.: Differential privacy: A survey of results. TAMC 2008: Theory and Applications of Models of Computation **4978 LNCS**, 1–19 (2008)
6. Friedler, S.A., Choudhary, S., Scheidegger, C., Hamilton, E.P., Venkatasubramanian, S., Roth, D.: A comparative study of fairness-enhancing interventions in machine learning. FAT* 2019 (2019)
7. Helberger, N., Pierson, J., Poell, T.: Governing online platforms: From contested to cooperative responsibility. The Information Society **34**(1), 1–14 (2018)
8. Hesselman, C., Grosso, P., Holz, R., Kuipers, F., Xue, J.H., Jonker, M., de Ruiter, J., Sperotto, A., van Rijswijk-Deij, R., Moura, G.C., Pras, A., de Laat, C.: A Responsible Internet to Increase Trust in the Digital World. Journal of Network and Systems Management **28**(4), 882–922 (2020)
9. Kamiran, F., Žliobaitė, I., Calders, T.: Quantifying explainable discrimination and removing illegal discrimination in automated decision making. Knowledge and Information Systems **35**(3), 613–644 (2013)
10. Nissenbaum, H.: Privacy In Context: Technology Policy And The Integrity Of Social Life. Stanford Law Books, Stanford University Press (2009)
11. Pearl, J.: Understanding Simpson's Paradox. The American Statistician **68**(1), 8–13 (2014)
12. Rieder, B., Gordon, G., Sileno, G.: Mapping value(s) in AI: the case of YouTube. In: AoIR 2020: The 21th Annual Conference of the Association of Internet Researchers (2020)
13. Saillenfest, A., Dessalles, J.L.: Role of Kolmogorov Complexity on Interest in Moral Dilemma Stories. In: Proceedings of the 34th Annual Conference of the Cognitive Science Society. pp. 947–952 (2012)
14. Sileno, G., Boer, A., van Engers, T.: The Role of Normware in Trustworthy and Explainable AI. In: 1st XAILA workshop on eXplainable AI and Law, in conjunction with JURIX 2018 (2018)
15. Sileno, G., Saillenfest, A., Dessalles, J.L.: A Computational Model of Moral and Legal Responsibility via Simplicity Theory. JURIX 2017 **FAIA 302**, 171–176 (2017)
16. Wachter, S., Mittelstadt, B., Russell, C.: Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI. SSRN Electronic Journal pp. 1–72 (2020)