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

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

DOI
10.1007/978-3-030-89811-3_14
Publication date
2021
Document Version
Final published version
Published in
AI Approaches to the Complexity of Legal Systems XI-XII
License
Article 25fa Dutch Copyright Act

Link to publication

Citation for published version (APA):
Sileno, G., Boer, A., Gordon, G., & Rieder, B. (2021). Like Circles in the Water: Responsibility as a System-Level Function. In V. Rodríguez-Doncel, M. Palmirani, M. Araszkiewicz, P. Casanovas, U. Pagallo, & G. Sartor (Eds.), *AI Approaches to the Complexity of Legal Systems XI-XII: AICOL International Workshops 2018 and 2020: AICOL-XI@JURIX 2018, AICOL-XII@JURIX 2020, XAILA@JURIX 2020 : revised selected papers* (pp. 198-211). (Lecture Notes in Computer Science; Vol. 13048), (Lecture Notes in Artificial Intelligence). Springer. https://doi.org/10.1007/978-3-030-89811-3_14

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
g.sileno@uva.nl
[2] KPMG, Amsterdam, The Netherlands

Abstract. What eventually determines the semantics of algorithmic decision-making is 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 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 (meaning and performativity) 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. For this reason, computational components need to be seen in

This research was partly supported by the UvA (RPA Human(e) AI seed grant) and by NWO (DL4LD project, no. 628.009.001).

V. Rodríguez-Doncel et al. (Eds.): AICOL-XI 2018/AICOL-XII 2020/XAILA 2020, LNAI 13048, pp. 198–211, 2021.
https://doi.org/10.1007/978-3-030-89811-3_14

an "ecological" perspective, i.e. whose working is entrenched with the operations of other computational modules, and whose deployment is driven-by and applies-on heterogeneous human social systems, characterized by competing interests, distinct socio-economic positions and possibly incompatible preferences.

In parallel work [20], we started exploring methods to investigate how "values" are generated, distributed, and translated between contextualized social processes and automatic/automated decision-making components; inspired by the idea of *encircling* introduced in security studies [4], 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 objective 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* [18], and *contextual demographic disparity* [28]. Section 3 elaborates on the function and functioning of *responsibility* as a cognitive mechanism, proposing the concept of "agentive responsibility". Section 4 considers a recent proposal on *responsible Internet* [12] revisiting the *accountability-responsibility-transparency* (ART) principles for AI [5] in the domain of networking, and elaborates on how to extend 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 the problem of selecting or producing adequate data. Following the overview given in [8], 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.

2.1 Computational Reflection

A relevant framework through which to look at these interventions is provided by the notion of *computational reflection*, i.e. the ability of a system to inspect and modify itself in order to improve its performance (see e.g. [2]), generally further distinguished in:

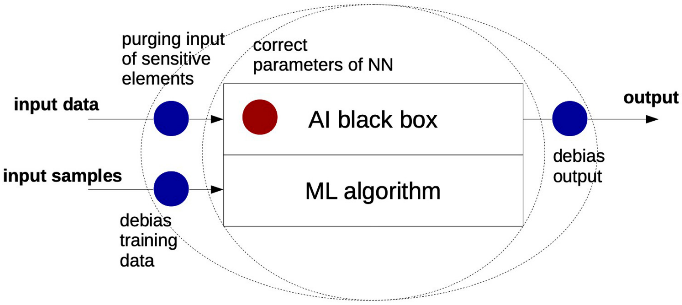

Fig. 1. Most common interventions proposed by algorithmic fairness solutions, illustrated in terms of computational reflection: behavioural (blue circles) or structural (red circle). (Color figure online)

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.

In order to be effective, behavioural reflection requires the system to be aware of its "semantics" with respect to its environment, i.e. of the processes it may initiate, inhibit or may be involved into, whereas structural reflection only needs the system to be able to look at its own components (data structures, procedures) and modify them according to some criteria.

Through the lens of computational reflection, interventions of the types (1), (2), (4) become examples of behavioural reflection: they introduce additional modules to process the input before the output and/or after the core module, without modifying it structurally; (3) is instead an example of structural reflection, concerning in particular the neural network parameters (see Fig. 1). In all these cases the focus is clearly on *data*, either input data, output data or data relative to the model. The knowledge used to guide behavioural reflection does not go beyond the qualification of which types of data are sensitive/protected.

2.2 From Data to Data-Flow

Methods based on behavioural reflection (e.g. the blue circles in Fig. 1) suggests that, alternatively, one can see fairness as a problem of *data-flow*: i.e. of intervening or constraining adequately the connections between the data processing components.

On a fundamental level, any computational component can seen as an assemblage of lower-level computational components. Even a Turing machine can be mapped to a functionally equivalent distributed system, whose individual components activate other components, performing in turn activating actions, and so on. In fact, parallel models of computation have been proven to be more general than traditional sequential/procedural models (e.g. [13]). Other computational models, like those applied for computer networks or for neural networks, can be directly looked through actor-based lenses. See for instance the recent introduction of *agnostic networks* [9], based on the intuition that training could be based not on modifying the weights of the network components, but on tinkering with their connections. This transforms our perspective on the machine learning problem from being *metrical* in nature (e.g. targeting an adequate latent space given a fixed network topology), to being explicitly *topological*.

This change of perspective facilitates the convergence of various problems into one of *responsible processing of informational flows*.[1] For instance, privacy can be seen as a set of limited rights and abilities controlling disclosure-of (i.e. channels transmitting) self-information. *Differential privacy* methods [7], introduced to protect against the reconstruction of data of individuals by intersection of a sufficient number of queries, work by adding a certain amount of noise to the output provided by a curator (that is, a module responding to queries), or by means of a stochastic curator. Both differential privacy templates can be functionally interpreted as an addition of external noise channels, destroying part of the information by interference. Opposite to protective measures, there exist initiatives and solutions that support or facilitate the construction of informational connections, e.g. as those driven by the FAIR data principles (*findable-accessible-interoperable-reusable*) [29], and, with distinct bases and purposes, the various Open Data initiatives (for governmental data, research data, etc.).

2.3 Responsible Computing as Responsible Disclosure

On a functional level, a data-flow perspective highlights the pivotal role of the *control of information disclosure*, which can be—in terms of information—*negative* (i.e. restricting, limiting disclosure) or *positive* (i.e. enabling, granting it). However, this qualification does not say anything about the practical (i.e. non-informational) effects of processing, or more precisely, of the use of such processing.

Let us consider a machine learning application used in support for decision-making, schematized in Fig. 2. Let us separate the computational module running the machine learning method from the module producing inferences by means of the model parameters extracted by the first. The sample data used for training typically follows a different routing than the data used for prediction; for instance, they come from two distinct data-providers, which in turn collect data belonging to different data-subjects. The inferential module can be used by several data-processors, in turn controlled by distinct data-users, that possibly utilize the output produced by the processors for their specific purposes. This derived information may intervene in the users' decision-making, determining *counterfactually* a certain decision resulting in positive or negative action. Depending on the actual environmental disposition, each individual decision (suppose e.g. about recruitment) would produce effects affecting not only

[1] We include in the term "flow" interface aspects: how outputs are presented, and how user-side interventions are able to act on processing.

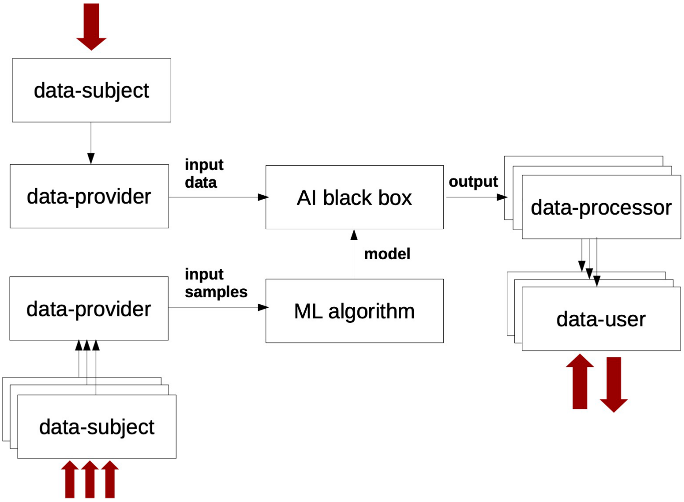

Fig. 2. Schematic data and value flows associated to training and use of a machine-learning application processing personal data in support to decision-making.

the data-user (e.g. the recruiter), but typically also the data-subject whose data was processed (e.g. the target candidate), and other parties (e.g. competitor candidates, the families of candidate and competitors, their communities, etc.). Potentially, it may indirectly affect also the data-subjects that provided the data for the training, depending on the ramification of the consequences.

This schematization serves two purposes here. First, it highlights that contemporary technological challenges are not only a matter of responsible machine learning, but rather of *responsible computing* (including processing, data-sharing, networking, etc.). Under this lens, privacy can be seen as a set of limited rights or abilities to control disclosure of self-information when this information has the potential to affect outcomes for the individual. Dually, excess of control on information disclosure opens up to abilities to obtain unfair outcomes, as (direct or indirect) *discrimination*, or even *fraudulent schemes*.[2] Second, removing the boundaries between components internal and external to the system, and looking at them as a network in which certain information flows (or does not flow) producing some (positive, negative) impact on the social participants, unveils that the main difference between structural and behavioural reflection concerns the amount of observability/controllability on the network required for the reflection to be applied. As soon as we consider humans or other artificial components producing and further processing that data, the depth required for

[2] Administrative, corporate and other types of frauds are typically conducted by exploiting an overlap of roles over the same identity (a physical person, an organization, etc.) [1,24]; such an overlap enables access to information that should be otherwise inaccessible.

a proper behavioural reflection naturally increases, creating figuratively “circles in the water” of components interfering/interacting with each other.

3 The Role of Context

At face value, technical solutions as those proposed for *algorithmic fairness* or *differential privacy* do not bring to the foreground that 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 [3]:

- *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 are reflected in distinct definitions that are incompatible amongst each other, and, furthermore, they can produce effects which are still detrimental to the protected classes [3]. Even at a technical level, it is recognized that something is missing in the picture. Indeed, the requirement of adding context in issues about privacy and discrimination is the starting point in several works looking at the problem from a higher-level (typically socio-legal) perspective. We will consider here two examples in this respect.

3.1 Contextual Integrity

The well-known framework of *contextual integrity* by Nissenbaum [18] 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 (e.g. interests of parties, societal norms and practices). 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. Context is defined not 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), which is used by the subject and other parties to form their expectations. The "ecological" nature of all these contextual elements makes it difficult if not impossible to capture them *monistically* within the informational artefacts which are target of directives about disclosure. In other words, context maps to layers well beyond the system boundaries, but still, it provides crucial terms to define the system "semantics" required for an appropriate behavioural reflection.

3.2 Contextual Demographic Disparity

Recent work by Wachter et al. [28] analyzes the concept of *contextual demographic (dis)parity* (CDD) (based on the measure of *conditional (non-)discrimination* proposed by Kamiran et al. in [15]), 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, caused by the diverse "composition of the disadvantaged and advantaged group, the severity and type of harm suffered, and requirements for the relevance and admissibility of evidence". They 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. Rephrased in the terms of behavioural reflection: algorithmic-driven assessment can explore a larger coverage of the network, but further layers exist beyond that, requiring human experts to stay in the decision-making loop.

Let us have a further look at CDD. Suppose a norm aims to protect certain groups of people (identified by means of protected attributes), and suppose a certain decision process produces a positive or negative outcome (according to a certain value structure), 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, where R are conditions used to divide the population into sub-populations, A_R is the proportion of people of the sub-population with protected attributes put in the advantaged class, D_R is the proportion with protected attributes put in the disadvantaged class.

But how to decide upon the possible R to take into account? Following Kamiran [15], 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 [19], 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 need to be added to the picture in order to evaluate the "reverberations" of the agents' actions throughout the system.

4 Function and Types of Responsibility

Human communities exhibit ascription of responsibility as a spontaneous, seemingly universal behavior, but variants of this construct can be found even in technical and abstract domains. Here we propose a simplified outline, with no pretension of being exhaustive, or covering all the available views on those matters. Our aim is to highlight some specific characteristics that are generally overlooked in technical settings.

4.1 Function of Responsibility

From a systematic standpoint, 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. Software engineers, for instance, are suggested to follow e.g. the *single-responsibility* design principle (one module encapsulates one functionality)—one of the SOLID principles for object-oriented programming, and also referred to in *agile development* [17]—because it helps to localize bugs.

4.2 Epistemic Responsibility

Practical failures directly map to failures of *expectations*, namely with respect to the mechanisms attributed to the system or its components. On a conceptual level, those mechanisms, and the events fed to them as inputs, contribute to our understanding of the system as a whole, and for this reason they can be given *epistemic responsibility*. Also in this case, lack of predictive power exhibited at the occurrence of a failure triggers remedy/repair functions, i.e. an investigation on whether the provided input is correct, or the search of better mechanisms to be assigned, or constructed if not available. In stratified systems such *retrodictive*, *explanatory* construction might be a recursive process, targeting defective lower-level components. However, a similar analysis can be also executed in absence of failure, to understand on which basis the system works. Indeed, *explainable AI* techniques leverage concepts as e.g. *Shapley value* (SHAP [16])—the payoff a player can expect in a coalition game for its contribution to the outcome—to interpret the contribution of a certain feature in producing a certain conclusion.

4.3 Causal Responsibility

Returning on the practical dimension, *causal responsibility* is meant to identify which ones, amongst the components involved in a chain of events, *actually caused* (or prevented) a certain outcome. Several properties have been identified (not without discussion) in the literature related to actual causation, as e.g. *counterfactuality* (the outcome would not have occurred if that agent had behaved otherwise), and *sufficiency* (agent behaviour was the ultimate determinant of the outcome). In general, however, some degree of responsibility is also assigned to *concurrent contributions*, i.e. to enabling conditions that allowed a sufficient event to occur; a problem associated to this extension is to identify the relative contribution of causes (see e.g. the experiments in [26]).

4.4 Moral Responsibility

Moral responsibility builds upon causal responsibility (although in some circumstances it might overdeterminate it), but it also presupposes a preferential structure (or of an underlying value structure) about outcomes in the world. Although certain contributions in the analytical literature (e.g. [10]) neglect this aspect, *blame* or *praise* would not make sense for morally irrelevant outcomes.

Empirical studies (e.g. [22], for a unifying computational model see e.g. [27]) suggest that moral responsibility:

- may hold for actions merely initiating potential causes of an outcome;
- grows with the impact of the outcome;
- is diminished e.g. if the action is not under the (expected) control of the agent, or the outcome is (justifiably) not foreesable from her standpoint.

4.5 Agentive Responsibility

Rather than facing the question of what makes an agent a moral agent, we can more conservatively identify, considering the previous concepts, 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. For instance, a dedicated module (cf. an expert) is expected to have better controllability and foreseeability than a general purpose module (cf. a layman person). 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*.

4.6 Accountability, Liability

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 [5]: *accountability*: motivations for the decision-making (values, norms, etc.) needs 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. Or the two requirements for *meaningful* (human) *control* [23]: *tracing*: the system needs to be able to trace back the outcome of its operations to specific directives given by humans during design or operational phases; *tracking*: the system needs to respond to (moral) reasons deemed relevant by directives given by humans guiding the system and to relevant facts in the environment in which the system operates. Or still, the seven requirements (human oversight, technical robustness, privacy and data governance, transparency, fairness, well-being, and accountability) identified by the expert group appointed by the European Commission [14].

At the moment, 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 and the computational domain as separated from the human domain.

5.1 Responsible Networking

Interestingly, a recent paper by Hesselman et al. on the concept of *responsible Internet* [12] 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 of Sect. 2): “the Internet has only one high-level task, which is to securely and reliably provide end-to-end communications”. This task needs to be solved on a decentralized architecture with distributed ownership and control.

The paper revisits and slightly modifies the ART principles [5], operationalizing 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.); *data accountability* holds if network operators explain the processing of specific data flows, e.g. their routing decisions or incidents during transmission; *infrastructure accountability* means that network operators explain their infrastructural design decisions. 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] Note that for implementing accountability, norms on which decisions are based need to be made explicit.

The authors also sketch an architecture of how a responsible Internet could work, consisting of three main components: a *network inspection plane* (NIP), enabling users to query the infrastructure for details about its internal operations in terms of network operators; a *network control plane* (NCP), enabling users to specify their expectations on the data which is transmitted by network operators, based on network descriptions; a *policy framework* (POL), enabling infrastructure maintainers to specify policies and have network operators abiding to certain norms, by means of auditing or other enforcement techniques.

5.2 From Operational to Agentive Responsibility?

How does this more operational view on responsibility relate to 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 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 (part of) 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.

An important comment on this point: in many aspects the term "risk" has already a prominent role in governance technology. However, as several authors

[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).

[4] Similar considerations apply looking beyond the technological boundaries, cf. Helberger et al. [11] 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.

observed (e.g. Rouvroy [21], Dillon [6]), 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. The account proposed here takes indeed this direction.

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* [25]), 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 and (plural) knowledge constructed from system practices, whether a certain requested processing is indeed justified. 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 has responsibility that the processing requested by the data-controller is complying with the rules.

References

1. Boer, A., van Engers, T.: An agent-based legal knowledge acquisition methodology for agile public administration. In: Proceedings of the 13th International Conference on Artificial Intelligence and Law (ICAIL 2011), pp. 171–180. ACM, New York (2011)
2. Capra, L., Blair, G.S., Mascolo, C., Emmerich, W., Grace, P.: Exploiting reflection in mobile computing middleware. ACM SIGMOBILE Mob. Comput. Commun. Rev. **6**(4), 34–44 (2002)
3. Corbett-Davies, S., Goel, S.: The measure and mismeasure of fairness: a critical review of fair machine learning (2018)
4. De Goede, M., Bosma, E., Pallister-Wilkins, P.: Secrecy and Methods in Security Research: A Guide to Qualitative Fieldwork. Routledge, New York (2019)
5. Dignum, V.: Responsible autonomy. In: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI), pp. 4698–4704 (2017)
6. Dillon, M.: Underwriting security. Sec. Dialogue **39**(2–3), 309–332 (2008)

7. Dwork, C.: Differential privacy: a survey of results. In: Agrawal, M., Du, D., Duan, Z., Li, A. (eds.) TAMC 2008. LNCS, vol. 4978, pp. 1–19. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-79228-4_1
8. Friedler, S.A., Choudhary, S., Scheidegger, C., Hamilton, E.P., Venkatasubramanian, S., Roth, D.: A comparative study of fairness-enhancing interventions in machine learning. In: FAT* 2019 (2019)
9. Gaier, A., Ha, D.: Weight agnostic neural networks. Adv. Neural Inf. Process. Syst. **32**(NeurIPS), 1–19 (2019)
10. Halpern, J.Y.: Cause, responsibility and blame: a structural-model approach. Law, Prob. Risk **14**(2), 91–118 (2015)
11. Helberger, N., Pierson, J., Poell, T.: Governing online platforms: from contested to cooperative responsibility. Inf. Soc. **34**(1), 1–14 (2018)
12. Hesselman, C., et al.: A responsible internet to increase trust in the digital world. J. Netw. Syst. Manag. **28**(4), 882–922 (2020)
13. Hewitt, C.: What is computation? Actor model versus turing's model. In: A Computable Universe: Understanding and Exploring Nature as Computation, pp. 159–186 (2012)
14. High-Level Expert Group on Artificial Intelligence (AI HLEG): Ethics Guidelines for Trustworthy AI (2019)
15. Kamiran, F., Žliobaitė, I., Calders, T.: Quantifying explainable discrimination and removing illegal discrimination in automated decision making. Knowl. Inf. Syst. **35**(3), 613–644 (2013)
16. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proceedings of Advances in Neural Information Processing Systems (NIPS), pp. 4766–4775 (2017)
17. Martin, R., Rabaey, J., Chandrakasan, A., Nikolic, B.: Agile Software Development: Principles, Patterns, and Practices. Pearson Education, New Jersey (2003)
18. Nissenbaum, H.: Privacy In Context: Technology Policy And The Integrity Of Social Life. Stanford University Press, Stanford (2009)
19. Pearl, J.: Understanding Simpson's Paradox. Am. Stat. **68**(1), 8–13 (2014)
20. 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)
21. Rouvroy, A.: The end(s) of critique. Priv. Due Process. Comput. Turn **39**, 143–167 (2008)
22. 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)
23. Santoni de Sio, F., van den Hoven, J.: Meaningful human control over autonomous systems: a philosophical account. Front. Rob. AI **5**, 15 (2018)
24. Sileno, G., Boer, A., van Engers, T.: Reading agendas between the lines, an exercise. Artif. Intell. Law **25**(1), 89–106 (2017)
25. 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)
26. Sileno, G., Dessalles, J.L.: Qualifying Causes as Pertinent. In: Proceedings of the 40th conference of the Cognitive Science Society (CogSci 2018), vol. 1, no. 2 (2018)
27. Sileno, G., Saillenfest, A., Dessalles, J.L.: A Computational Model of Moral and Legal Responsibility via Simplicity Theory. JURIX 2017 FAIA 302, pp. 171–176, Amsterdam (2017)

28. Wachter, S., Mittelstadt, B., Russell, C.: Why fairness cannot be automated: bridging the gap between EU non-discrimination law and AI. In: SSRN Electronic Journal, pp. 1–72 (2020)
29. Wilkinson, M.D., et al.: The FAIR guiding principles for scientific data management and stewardship. Sci. Data **3**, 1–9 (2016)