A discussion frame for explaining records that are based on algorithmic output

INTRODUCTION

Records provide evidence of past activities. If someone for instance is denied a parking permit meant for disabled persons, because the requirements for obtaining it are not met, there should be an intelligible record stating the refusal. The record will probably be kept for some time for further inspection, or for handling of complaints. For the sake of transparency and accountability, the content of the record of course must be understandable. Further, the contextual information and the recordkeeping practices are essential preconditions for trusting this evidence. Sound information and records management provide transparency and enable accountability (see for instance Willis, 2005). These are bread-and-butter assumptions of the records management discipline.

Records that are output from automated or algorithmic processes do not necessarily differ much from manually created, captured and organized records in matters of evidential value and trustworthiness. The contextual information still testifies to what activities the records arose from. However, explaining the content of such records may be more difficult. A human who answers an e-mail could explain a verbose line of reasoning, or he could invoke a company procedure he has followed in order to explain the matter in question. In the first case, the explanation originates from the process the e-mail was a part of. In the second case, it is referenced in an indirect way. The content of the company procedure providing the explanation originated from a separate process, outside the process the e-mail belongs to. If the content of a record is the output from an automated decision, or even from a fuzzy data-imbued algorithm, an explanation might be offered in a similar fashion, as a postscript, external to the operational activities. In complex originating situations, the content of the record can be difficult to explain, trace or recalculate after the fact. An extreme example would be autonomous weapon systems, picking targets for bombing based on analysis of perceived actions by an enemy, without a human commander pushing a button to fire the arms. Accountability and transparency will itself be a complex issue, regardless of what is recorded (Liu, 2016). A record that merely states what target the weapon system picked, is not enough to analyse what actually happened. No one can explain the actions of the system unless both the information available and the inferences drawn from it is captured. Any hope of analysing what happened relies on what is recorded. Sufficient explanation cannot be obtained from studying process flow or computer program code alone.

This paper explores some concepts from the records management field, from the theory of science, and from legal theory, attempting to establish a discussion frame for explaining the content of records that are based on an algorithmic output. What is meant by a discussion frame in this paper, is a combination of a small number of concepts from different academic fields that could be suitable when exploring the needs for, and the possibilities of, such explanations. The contribution from the records management field has already been hinted at; information that is external to the recording process itself should sometimes be understood, and managed, as a different kind of record. A contribution to the discussion frame from the theory of science is the debate on similarities and differences between explanations and predictions. For an algorithm with an uncertain outcome, the least unlikely prediction could be the closest available approximation to an explanation. The third contribution to the discussion frame is from legal theory, regarding whether there are legal obligations to provide explanations, or a legal right to obtain an explanation. In the field of data protection law, there is some debate on the individuals' right to an explanation for automated decisions. A right to an explanation would also imply a reciprocal obligation for someone to provide the explanation. The right to an explanation takes on different nuances due to whether it is interpreted as a general need for system transparency, or as an obligation to explain what gave rise to the content of specific information or records. The legal obligation to provide an explanation could also take alternative paths, such as a unilateral obligation to undertake ethical assessments for algorithms, or an obligation to carry out risk assessments, as a means to curb incomprehensible and possibly harmful algorithms, without necessarily granting rights to individuals who are affected by them. In this paper, the algorithm ethics approach is put forward mainly as an alternative legal approach. However, there is also a growing literature on algorithm ethics in other fields discussing ethical implications of algorithms' effects on trust in media or other societal institutions (Diakopoulos, 2016; Diakopoulos 2014).

POLICY TRANSACTIONS, OR 'DOCUMENTED INFORMATION'

Large numbers of records offer an explanation about what they are, and why they were created, due to familiar appearances or genre features. For instance, if an application is part of an administrative process, both the record metadata, visual elements and body text will easily reveal which records contain completed application forms. On the other hand, each completed application form will probably not explain why it included the chosen selection of information elements. Understanding the content of the application form, or making a judgement of whether the information was correct and sufficient, requires knowledge of the business process it was a part of. Knowledge of the process is in most cases external to the records arising from it.

Theodore Schellenberg, in his 1956 volume 'Modern Archives', drew a distinction between policy transactions on the one hand, and operational transactions on the other, both emanating from functions and activities (Schellenberg, 2003, p. 39, also shown in an illustration on p. 55):

> *Records on policy and procedural matters—on general as distinct from specific matters—are difficult to assemble, to organize into recognizable file units, and to identify in such a way that their significance will be made known. Records of routine operations, on the other hand, are easily classified.*
> *Important records are difficult to retire after their current uses have been exhausted. Important records on policy and procedure do not become obsolete, or noncurrent, as soon as the transactions in connection with which they may have been made are completed. The policies and procedures they establish often continue in effect. And even if those policies and procedures are superseded, the records of them serve to explain and give meaning to the change.*

As Schellenberg's general view on records is that they are 'only a byproduct of administrative activity' (p. 46), it follows that the policy and procedure activities themselves are important activities generating their own vital records. The records pertaining to 'general matters' remain current after the activity they originate from is completed, they can and should be invoked whenever relevant to the operational activities they apply to. The point in bringing Schellenberg's concept of policy transactions into this discussion is to show that records governing other records is a long-standing aspect of records management. A lack of relevant policy records may decrease the ability to understand and explain operational records.

More recently, the need to manage and secure records on policy and procedures has been highlighted in the family of Management System Standards, abbreviated MSS, from The International Organization for Standardization (ISO). Standards within this family prescribe a cyclic procedure of planning, establishing, following and verifying policies and procedures to achieve a management objective. Different MSS standards apply to different objectives, such as the ISO 9000-series for quality management, the ISO 14000-series for environmental management, and the ISO 27000-series for information security management. From 2011, the records management discipline joined the MSS family of standards when the ISO 30300-series, management system for records, sometimes abbreviated MSR, was established. Over the last few years, ISO has made efforts to harmonize the template for different standards in the MSS family into a common high-level structure. The term 'documented information' has, since 2015, been the ISO MSS concept covering the policy, procedure and verification documentation (or records) associated with all phases of their cyclic management processes.

In the most recent version of the ISO standard on management systems for records, it is explicitly stated that '[d]ocumented information of the MSR is part of the records of an organization, which shall be managed in a records system.' (ISO 30301:2019, clause 7.5.3, p. 10). This includes some modest requirements for identification of such documentation, and various metadata to establish authorization, versioning and retention control for the documented information.

The term 'documented information' in the ISO MSS family of standards resembles Schellenberg's records emanating from policy transaction, but also takes it a bit further by including documentation related to verification and evaluation of the management procedures.

This distinction between two kinds of records, 'records on policy and procedures' on one hand, and records on each specific instance of business activity on the other, forms the part of the discussion frame that is contributed from the field of records management. In this paper, inspired by Schellenberg and for the sake of simplicity, these two kinds of records are labelled policy records and operational records. An explanation that helps understanding algorithmic output may reside in either kind. However, these two kinds of records seem to bring about different qualities to the explanations they embody.

EXPLANATION VERSUS PREDICTION

A minimum expectation for a record is intelligible content, making it possible to comprehend what it states. Almost equally important is the trustworthiness of the record, involving reliable records management processes, in a broad sense. The aim of this paper is to enable discussion on how records may make algorithmic output understandable. The colloquial term for what the records should provide then, is an *explanation*. An explanation entails an ability to convey how and why the state of affairs reflected in the records came about.

Common and widespread examples of records resulting from algorithmic output are those created by conventional automated public sector case handling systems. One out of many examples from Norwegian authorities is the administration of students' loans and grants. The rules implemented in the system let students apply for a loan, or a grant, or both. Students who live with their parents may get a loan, but they will not get a grant. Students who do not live with their parents may get both a loan and a grant, depending on what they applied for. The case handling system receives signed digital application forms, gathers some more data from various sources, and starts churning its fixed repertoire of criteria, rules and amounts. Knowing the rules, it will be fairly straightforward to explain each decision resulting from this system. For this kind of deterministic and orderly algorithmic output, an everyday language understanding of the term ‘explanation’ will suffice.

In the new era of algorithms, involving *inter alia* large amounts of volatile data, machine learning, or probabilistic outcomes, it may be harder to express an explanation in terms of a linear path from criteria to outcomes. In some situations, this is merely a pedagogical problem. The linear path of how and why is still there, but it can be difficult to present it in a way that average humans can understand. However, some situations will also occur where different outcomes or competing explanations are possible. Documenting the algorithmic output could for instance be achieved by representing a range of possible explanations, or by picking the explanation that most likely has determined the outcome. Both of these strategies for explaining complex or cluttered algorithmic output involve an element of prediction. To explore the relationship between explanation and prediction, it might be useful to revisit briefly a debate on this topic in the philosophy of science from the mid-twentieth century.

A position developed by Carl Hempel and fellow proponents of so-called logical positivism, was to regard predictions as essentially the same as explanations. ‘[T]he logical structure of a scientific prediction is the same as that of a scientific explanation’ (Hempel, 1942, p. 38). The underlying premise was that an explanation involved two components, the first one is a known initial condition, the second is a ‘covering law’, a general scientific law or a comparable causal mechanism with a high degree of regularity. A prediction involves the same two components and is therefore the same. The difference between explanation and prediction, according to this position, is merely pragmatic - in an explanation the final event is known. A problem with this position is the reliance on a ‘covering law’, restricting the explanations (and by symmetry, the predictions) to be valid only for algorithms operating on a controllable information environment, thus missing out the ability to provide adequate explanations in the new era of algorithms.

The view that predictions and explanations are essentially the same gained some ground, and has also been adopted in situations where the ‘covering law’ is more dubious. For instance, as a way to understand indeterminism in quantum mechanism. Karl Popper stated in an article in two parts that ‘explanation, in the scientific sense, is the same as prediction, except that the demand is dropped that the deduced statement must be obtained *before* the event which it describes’ (Popper, 1950 p. 191).

In the last part of the 1950s and onwards, several contributions to the philosophy of science denounced the position that explanations and predictions were the same. A distinguishing feature is that explanations are certain, while a prediction only needs to be more likely than other alternatives. This point was made by Nicholas Rescher in 1958, where he specifically pointed out a difference in the need for justification of the predictions.

> *The thesis I wish to stress is that the reasoned validation of a prediction—the presentation of reasoned justifying arguments in support of the prediction—need do no more than render its conclusion significantly more likely than its principal alternatives. In this there resides a crucial difference between predictions and explanations. An adequate explanation must render its conclusion virtually certain, and thus tenable* per se*, while a soundly reasoned prediction need do no more than render its conclusion relatively tenable, i.e. more tenable than alternatives, and to do this in such a way that a* sufficient *(rather than* conclusive*) reason is forthcoming for espousal of the predicted eventuality in preference to other possibilities* (Rescher, 1958 p. 286).

An explanation serves to defend a result. Before a prediction can be used for defending a result, it is itself in need of defence, a justification for why it is more likely than its alternatives. However, explanations are not necessarily easier to understand than predictions, nor is it the other way around. The pedagogical issue of an understandable presentation is bracketed off for now.

The view that predictions are different from explanations eventually 'won' the philosophy of science debate of the mid-twentieth century. According to Heather Douglas, this led to an almost complete abandonment of prediction in discussions of explanation, which in turn 'has hampered our ability to properly understand explanation' (Douglas, 2009 p. 445). Although she grants that explanations and predictions are not the same, there is an important relationship between them that should be reintroduced into the discourse. The proposed relationship 'is a tight, functional one: explanations provide the cognitive path to predictions, which then serve to test and refine the explanations' (Douglas, 2009 p. 454).

A feature of this relationship between explanation and prediction is that the threshold for accepting the likelihood of a prediction can in principle be lower than the threshold for accepting an explanation's claim to truth or certainty. As a corollary, it is sometimes possible to predict an outcome, and accept the prediction as justified, without being able to explain the event after it has occurred. A similar tangle has been pointed out by David Schum who mentions some aspects of child psychiatry as an example of situations where one can predict outcomes that one cannot explain (Schum, 2001 p. 198).

In some machine learning algorithms, an explanation in the sense of a certain, or 'true', account of the path from criteria to outcome will not necessarily exist. The algorithm may tweak its own parameters for interpreting data based on the successfulness of its predictions. Strategies for understanding and verifying machine learning algorithms are often based on interpretation and justification of the predictions (Biran & Cotton, 2017). Justification of the predictions can typically be to verify a small sample of known possible outcomes against the algorithmic output.

It may be perceived as unsatisfactory to rely on the least unlikely predictions in order to understand results, in lieu of unequivocal explanation, when the latter is hard to obtain. However, the relationship between explanation and prediction does at least offer a way into this troubled area. Building on the position that explanations and predictions are closely related, but differ in the degree of certainty, explanations and predictions can be viewed as different positions on an axis. Thus, the term 'explanation' in this discussion frame will mean information provided in order to shed light on how and why the state of affairs reflected in the records came about. This means explanations for records that are based on algorithmic output will differ, depending on what degree of certainty and justifiability is achievable.

ALGORITHMS AND THEIR DATA ENVIRONMENT

An algorithm is, broadly, a step-by-step procedure for solving a problem or accomplishing a goal, in a finite number of steps. In most practical situations, an algorithm is deterministic, it will produce the same output from a given input. It could be compared to a cooking recipe. Following a recipe where both the method and the ingredients are the same, will give the same food as a result. Changing the ingredients gives a different result. Algorithms can be conceived of as repeatable and reliable. They can still be hard to explain and understand, if the steps themselves or the branching and looping of the steps are complex.

In a new era of algorithms, the intuitive notion is that algorithms do more than running some operations where a known input yields the expected results. Large amounts of volatile data, machine learning, targeted recommendations, predictions of behaviour and other emerging applications of algorithms have entered public discourse with a renewed force over the last few years. There is a growing literature on the concerns over bias and lack of transparency in modern algorithms. An early and much cited article on bias in algorithms makes a distinction between bias in the technology, bias in the social institutions the technology operates within, and emergent bias that may evolve at a later stage, through the use of the algorithms (Friedman & Nissenbaum, 1996). This third kind of bias, emergent bias, is of particular interest. In more recent literature it is sometimes characterized as filtering, or filter-bubbles (Bozdag, 2013). The filtering is a complex mesh of responding to user interactions and various known or unknown actants propagating data that algorithms and platforms may make use of.

It could be tempting to label such algorithms 'non-deterministic', but in computer science, the term 'non-deterministic algorithms' is used for algorithms where different runs on the same input can produce different output. The combination of complex algorithms and opaquely filtered data is not normally non-deterministic computing in this sense, it is more aptly understood as deterministic algorithms working on input where the enterprise that use or benefit from the algorithm exercise less control over the data environment. Conventional automated decisions, which may be explained with a high degree of certainty, will be based on a data environment where the enterprise control the meaning, scope and supply of data. Powerful algorithms that make advanced predictions or learn from the successfulness of their predictions are more likely to be based on volatile data, where there is a lower degree of control over intended meanings and scope of the data. There are, of course, varying degrees and nuances of controlled data environments. In the same manner as explanations and prediction may be viewed as different positions on an axis, varying degrees of control over the data environment can likewise also be thought of as different positions on an axis.

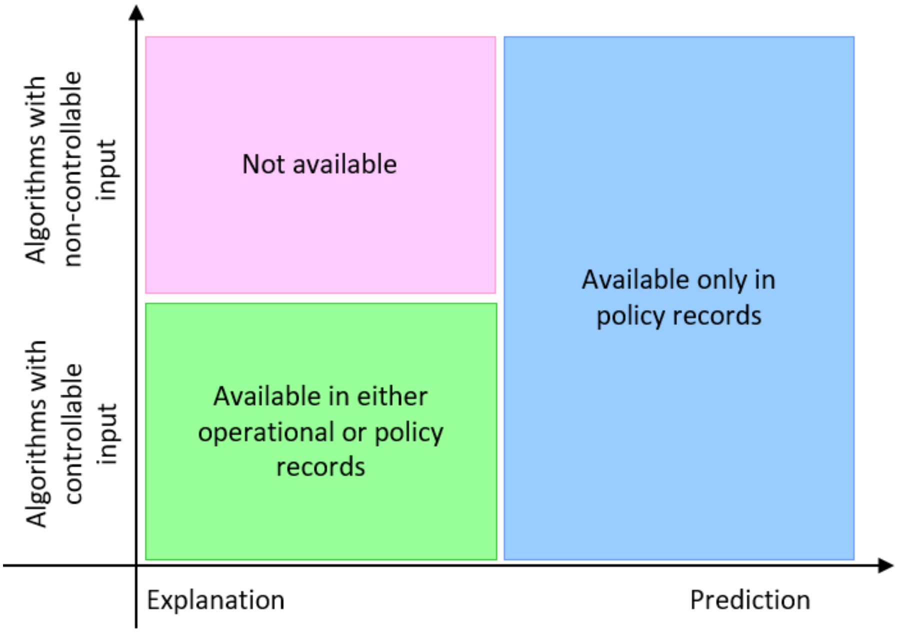

Algorithms with non-controllable input
Algorithms with controllable input
Not available
Available in either operational or policy records
Available only in policy records
Explanation
Prediction

Figure 1: Impact of degrees of control over input to algorithms

Figure 1 shows where explanations of algorithmic output in records may reside, illustrated in a two-dimensional plane. Explanations that both are certain, or 'true' after the event, and build on a highly controllable input, may be rendered in the operational records. However, such explanations can also be represented in policy records. For predictions that are likely but not certain, the explanations provided will only be available in policy records, regardless of whether the enterprise exercise a high or a low degree of control over the data environment. An explanation that is certain will neither be available in policy nor operational records, if the enterprise lack control over the data input.

Volatile data environments, algorithms and shifting boundaries between different kinds of actors, can reasonably be understood as phenomena – or assemblages of human and non-human actors – that straddle beyond the processing-information-technology nexus, and comprise societal institutions, norms and practices as well (Ananny, 2016). The complexities of the sociotechnical systems, which in the words of Ananny and Crawford (2018) is a complexity they enact, rather than contain, put a strain on a perceived direct inference both from transparency to understanding, and from transparency to accountability. They investigate a range of limitations to transparency as a means both to understand and to impose accountability. 'Holding an assemblage accountable requires not just seeing inside any one component of an assemblage but understanding how it works as a system' (Ananny and Crawford 2018, p. 983).

This paper thus puts the main emphasis on control over the data environment in order to differentiate various aspects of explaining algorithmic output in records. This does not by any means imply that it will always be easy to account for the actual steps and operations in an algorithm. One problem is to confine the algorithms. There is no universal method for referencing algorithms, or for telling exactly where a specific algorithm, that is supposed to be the unit to be explained, starts or ends within a system in operation. An algorithm as a phenomenon is a prescription, akin to a recipe; it may have different implementations. Sometimes algorithms are well known, or maybe even patented, and therefore suitable for being referenced. Quite often they are merely the abstract functioning behind an implemented program code. A quest for making algorithms more transparent, and to explain them to the satisfaction of those who need it, will probably in some instances take an effort of reification, turning abstract algorithms into 'things', so to speak. This could involve deciding where an algorithm start and ends, giving it a name, and express its functioning in some form of visualized flow or pseudocode. In a way, reification of this sort would be a step on the way to an explanation.

A second problem is that the algorithms are not necessarily implemented or residing in a system that the record-creating enterprise controls. The implementation of an algorithm could be in a 'cloud system' or other kind of outsourced environment, and it may be the intellectual property of another enterprise. In these situations, if the quest for transparency is taken seriously, the best available option is probably to make tenable explanations a part of what the enterprise buys from its vendor.

Brauneis and Goodman (2018) discuss legal aspects of controlling the data environment. They identify some principled impediments to algorithm transparency, including both the absence of appropriate record generation practices and privileges of confidentialty by government contractors. Their analysis of the problem of record practices is that '[m]any of the most important decisions in a big data application are made at the "wholesale" level of the design of a model, not at the "retail" level of application to a particular case' (Brauneis and Goodman, 2018, p. 153). Though the wording might be unfamiliar in the records management literature, wholesale and retail levels are metaphors that allude both to the same kind of problem that this paper refers to by the Schellenbergian distinction between policy records and operational records, and to understanding how the assemblage works as a system, as cited from Ananny and Crawford (2018) above. The primary suggestions Brauneis and Goodman make is for governments to use their contracting power to insist on appropriate record creation, provision and disclosure.

THE RIGHT TO AN EXPLANATION APPROACH IN GDPR

A wide variety of record creating situations involve processing of personal data. In the European Union, processing of personal data must comply with the General Data Protection Regulation, perhaps better known by the abbreviation GDPR (Regulation 2016/679/EU). GDPR includes a range of rules that aims to achieve transparency. A part of these rules that is of particular interest to understanding algorithms and their outcomes is the 'right to an explanation'. The GDPR does not literally express an unconditional right to an explanation, but the term explanation is mentioned as part of the 'suitable safeguards' that should be in place for automated decisions and profiling according to GDPR recital 71.

The basis for a right to explanation in GDPR is Article 22, which at the outset is 'a right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.' The remainder of Article 22 contains some exceptions from this right, as well as some obligations on the data controller to implement suitable measures to safeguard the data subject's interests. According to Article 22(3), the data subject will often have a right to contest such decisions.

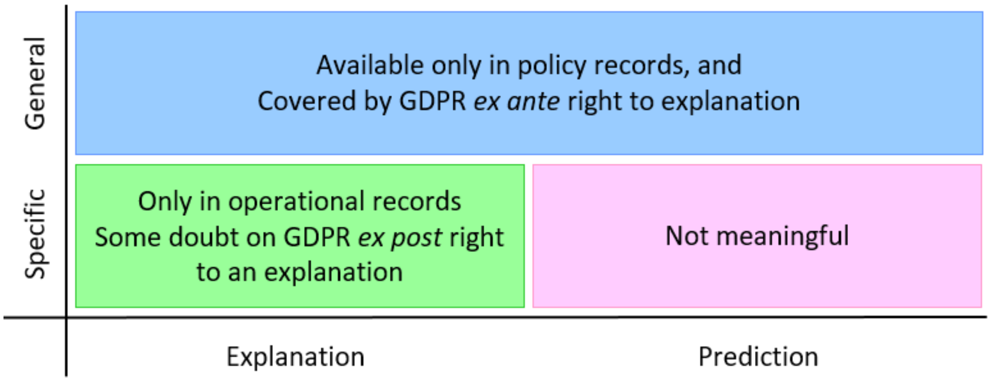
The right itself to obtain explanation is expressed in GDPR Articles 13(2)(f), 14(2)(g) and 15(1)(h). Article 13 concerns information received from the data subject, Article 14 concerns information received from others, while Article 15 concerns information given to the data subject on an access request. These three articles grant, in similar wordings, the data subject a right to information about 'the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.'

The expression 'meaningful information about the logic involved' is often referred to as a right to explanation in scholarly literature on data protection law. It Is worth noting that there is some debate within the data protection law community on whether, or to what extent, these GDPR articles really entail a right to explanation. In three different papers published within a short timeframe in 2017, the papers referring to each other, Selbst and Powles (2017) conclude that GDPR introduces an enhanced right to explanation, while this is disputed by Wachter, Mittelstadt, and Floridi (2017). The latter refine the argument of whether a right to an explanation would apply merely *ex ante*, or also *ex post*. A right to an explanation *ex ante* will be of general nature, limited to explaining system functionality, providing transparency concerning the intentions of the data controller. An *ex ante* right to meaningful information about the logic involved is a right to be informed about general features of the processing. A right to an explanation *ex post* would imply explanation of instantiated results of the automated processes, and their impact on the individual exercising his or her right. Wachter *et al* hold the view that the ambiguity and limited scope of Article 22 makes an *ex post* right to explanation infeasible. The third paper, Mendoza and Bygrave (2017), agrees partly on the views about ambiguity, but still concludes that an *ex post* right to explanation exists, especially emphasizing that it is a necessary precondition for the right to contest automated decisions as granted in Article 22(3).

The distinction between *ex ante* and *ex post* explanations in data protection theory resembles the distinction between policy records and operational records discussed above. This does not mean they are exact synonyms. *Ex ante* and *ex post* are primarily temporal categories. What the explanation may entail is associated with at what stage of the processing the explanation is obtained. Whether the explanation is limited to the functioning of the system, or it can be individualized as a trail of recorded actions, is in principle a side effect of what temporal category a supposed right belongs in. The distinction between policy records and operational records, on the other hand, refer to separate records creating processes. Policy records are not necessarily created before the operational records that they govern. The policy process might have been slow, or the policy records might have been superseded. The nature of an explanation, whether it is general or specific, is a matter of which kind of records the explanation resides in. It is not a matter of when the records that carry the explanation were created.

Explanations, as discussed above, can be of a more or less predictive nature. Viewed as different positions on an axis, explanations are more certain while predictions are justified by their likelihood. Any explanation, whether it is certain or if it is a likely prediction, can reside in the policy records. When the explanation resides in a policy record, it also has the quality of being a general explanation. An explanation in an operational record is, by definition, a specific explanation. It pertains to the specific event or transaction the record testifies to. A specific explanation, embodied in an operational record, will only be useful if it is understood as a certain, or 'true', explanation. An operational record, stating that 'this is a likely explanation of the algorithmic outcome', would probably cause confusion instead of enhanced transparency.

GDPR's expression 'meaningful information about the logic involved', is ambiguous in terms of whether it entails a right to a specific or a general explanation. The debate on a right to explanations *ex ante* versus *ex post* casst some doubt on whether a right to a specific explanation – embodied in operational records – is covered. It is less doubtful, however, that there is a right to a general explanation, although with some exceptions that is left out of this discussion. A second question of interest is whether GDPR's right to a general explanation is restricted to certain outcomes, or if it also includes likely outcomes, i.e. predictions. It is a fairly reasonable interpretation of the right to explanation, inferred from Articles 13, 14 and 15, to include explanations of a predictive nature. First, the expression 'meaningful information' provides some leeway on the exactness of the explanation. Further, the meaningful information also includes information about 'the envisaged consequences of such processing for the data subject'. An envisaged consequence could reach wider than a necessary consequence, and therefore include outcomes that are likely but not certain. And finally, the initial right not to be 'subject to a decision based solely on automated processing' in Article 22 also includes profiling as a form of automated processing. The term 'profiling' is defined in GDPR, Article 4(4), to mean '(…) use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements.' Hence, profiling comprises both 'certain' analyses and predictions.



	Explanation	Prediction
General	Available only in policy records, and Covered by GDPR *ex ante* right to explanation	
Specific	Only in operational records Some doubt on GDPR *ex post* right to an explanation	Not meaningful

Figure 2: Locus of general or specific explanations

The fourfold table in Figure 2 illustrates where specific or general explanations may reside, in terms of policy versus operational records and in terms of the GDPR ‘right to an explanation’.

ALTERNATIVE APPROACH, ALGORITHMIC ETHICS

Data protection provides an approach to transparency that combines rights of the individual and obligations on the responsible controller or processor. Other approaches that have been suggested in academic literature on algorithm transparency are to impose unilateral obligations to undertake ethical assessments or risk assessments before putting an algorithm into use. As algorithms can be value-laden, it would be reasonable to expect enterprises using the algorithms to be responsible for their implications (Martin, 2018). One imaginable kind of ethical obligation that would be very comprehensive in scope could be a regime of ethical approval, along the lines of authoritative Research Ethics Committees in clause 23 of the Helsinki Declaration on medical research ethics (World Medical Association, 2013/1964). However, a regime of external control of algorithms is probably not a viable option, in part because of issues with protection of trade secrets, and even more because the amounts of algorithms to assess would be overwhelming. The ethical principles enshrined in the Helsinki declaration could still provide an idea of what ethical assessments may achieve. *Inter alia*, medical research shall minimize risks of harm, the importance of the objective should outweigh risks and burdens on persons involved, and special considerations should be given to vulnerable groups or individuals. The principles and organizational requirements of the Helsinki Declaration provide guidelines that ultimately determine whether or not a medical research project can be initiated.

There is a growing literature on algorithmic ethics, a fairly broad account can be found for instance in Mittelstadt et al. (2016). A desire to increase the expectations on what enterprises ought to do is a consistent part of the algorithmic ethics discourse. Some emphasize empowerment of the data subjects, by way of consent or abilities to influence the input. Others propose that ethics should be about doing the right thing, regardless of involvement from the data subjects. This could take the form of a principle of minimizing harm by drawing a line on what an enterprise should allow algorithms to do in the first place, or an obligation to monitor emergent harms or bias after implementing the algorithm. While the individual rights approach is mainly discussed in legal literature, the literature on algorithmic ethics has emerged from a more diverse range of research communities, and generally appears more explorative and inquisitive. A particularly interesting side of the literature on algorithmic ethics is the deep recognition of the 'wholesale level' that is important to understand the outcomes (Brauneis and Goodman, 2018; Ananny and Crawford 2018).

Unilateral obligations on the enterprises to make ethical assessments, and further to monitor the effects of algorithms in use, also resembles the cyclic management systems propagated by the ISO MSS family of standards mentioned earlier in this paper. The cycles of a management system are based on risk assessments and mitigation, which might lack a portion of the 'niceness' often attributed to an ethics approach, but is otherwise essentially the same approach. The enterprise will need to ask themselves if they know what they are doing, if they are doing the right things, if they are doing it right, how sure they need to be, and to monitor whether intended or unintended effects, either caused by the enterprise itself or happening in their external environment, indicate a need to change anything. And then, to keep the cycle going. The advantage of branding this approach risk management, is that it plays into management processes that many enterprises, at least the larger ones, are familiar with and already have established in various areas. On the other hand, an incitement to assess the ethics will make the call for transparency, minimization of bias and avoiding burdens on vulnerable people clearer. The term algorithmic ethics may make a commitment to the outside world more visible than a wishy-washy 'reputational risk' that often represents external considerations in a risk management ethos.

Whether they are called risk assessments or ethics assessments, the important part in this context is that assessments are functions. They need to be part of an enterprise's machinery for making decisions on the use of algorithms, monitoring their effects, and reassessing whenever necessary. With the terminology introduced earlier, explanations and predictions will primarily be input to the assessment, as a means for the enterprise to understand what they are about to assess. It is necessary to understand what an algorithm will do, in order to make assessments. The documentation of the assessments will form policy records, that one might expect would embody the understanding of the algorithms that went into the assessments.

This alternative legal approach, algorithmic ethics, appears promising for achieving conscious and well-considered decisions by enterprises using algorithms. However, there can be a pitfall in how well suited the assessments will be as explanations. The understanding of an algorithm that the enterprise bases its assessment on is an understanding that is developed for the purpose of the assessment, making it more indirect or remote from the actual functioning of the algorithm compared to an explanation that is more explicitly formed to fit the purpose of providing transparency. In other words, linking explanations to the assessment process could lead to slightly tainted explanations. On the other hand, a cyclic assessment process secures regular reviews of the assessments, and consequently reviews of the explanations.

CONCLUSION

If the ambition of records management is to contribute to transparency and accountability, the records management community needs to be attentive to evolving demands on what records convey about the activities they represent. Both GDPR, requiring to some extent that an explanation for automated decisions and profiling can be obtained, and the emerging discourse on algorithmic ethics, point to societal expectations of increased transparency about what goes on inside the 'black boxes' of modern algorithms of various sorts.

This paper proposes a discussion frame for explanations of records that are based on algorithmic output. The elements of this discussion frame are a combination of a few concepts from different academic fields. A distinction between policy records and operational records is taken from the field of records management. A view of explanations and predictions as different positions on an axis, differing in degrees of certainty, is a contribution from the theory of science. Legal rights or obligations concerning explanations is the third part of the discussion frame. Such rights or obligations could be imposed by different kinds of laws, the most prominent examples in legal literature on algorithms have been related to data protection law, or to propositions on imposing legal obligations to undertake ethical assessments of algorithms.

The contribution of the records management field can certainly not be to provide tenable explanations for whatever algorithms might do. Explanations for the purpose of understanding algorithmic outcomes will, as with conventional records practices, require careful selections and interpretations regarding what contextual information is captured, and how. One contribution, which should be close at hand, building on the discussion in this paper, would be to point out what kinds of records, from what kinds of processes, explanations and predictions may reside in. However, the conundrum of transparency and understanding in complex environments may put some strain on parts of the basic concepts of records management. The characteristics of policy records, somewhat neatly confined by Schellenberg (2003) as important records created in a separate process, concerning general matters, and which according to ISO 30301:2019 themselves should be treated as records in a records system, were developed under a controlled environment assumption. If tenable and useful explanations require access to and interpretations of a plethora of internal and external influences on algorithmic outcomes, other conceptions of policy records might be needed. Records management research could pick different directions in addressing these matters, be it ‘boundary work’ that limits the scope of what records management should do, or applying known concepts for new situations, or seeking new ground and redefining parts of the established concepts in the field.

REFERENCES

Ananny, M. (2016). Toward an ethics of algorithms: Convening, observation, probability, and timeliness. *Science, Technology, & Human Values, 41*(1), pp. 93-117.

Ananny, M. and Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society, 20*(3), pp. 973-989. doi:10.1177/1461444816676645.

Biran, O. and Cotton, C. (2017), *Explanation and justification in machine learning: A survey.* Paper presented at the IJCAI-17 workshop on explainable AI (XAI).

Bozdag, E. (2013), Bias in algorithmic filtering and personalization. *Ethics and Information Technology, 15*(3), pp. 209-227. doi:10.1007/s10676-013-9321-6.

Brauneis, R. and Goodman, E. P. (2018). Algorithmic transparency for the smart city. *The Yale Journal of Law & Technology, 20*, pp. 103-176.

Diakopoulos, N. (2014). *Algorithmic accountability reporting: On the investigation of black boxes*. Columbia Journalism School, Tow Center for Digital Journalism.

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM, 59*(2), pp. 56-62.

Douglas, H. E. (2009), Reintroducing prediction to explanation. *Philosophy of Science, 76*(4), pp. 444-463.

Friedman, B. and Nissenbaum, H. (1996), Bias in computer systems. *ACM Transactions on Information Systems, 14*(3), pp. 330-347. doi:10.1145/230538.230561.

Hempel, C. G. (1942), The Function of General Laws in History. *The Journal of Philosophy, 39*(2), pp. 35-48. doi:10.2307/2017635.

ISO 30301:2019. *Information and documentation – Management systems for records – Requirements*. Geneva: International Organization for Standardization.

Liu, H.-Y. (2016), Refining Responsibility: Differentiating Two Types of Responsibility Issues Raised by Autonomous Weapons Systems. In N. Bhuta, S. Beck, R. Geiss, & H.-Y. Liu (Eds.), *Autonomous Weapons Systems: Law, Ethics, Policy*. Cambridge University Press, pp. 325-344.

Martin, K. (2018), Ethical Implications and Accountability of Algorithms. *Journal of Business Ethics*. doi:10.1007/s10551-018-3921-3.

Mendoza, I. and Bygrave, L. A. (2017), The Right Not to be Subject to Automated Decisions Based on Profiling. In T.-E. Synodinou, P. Jougleux, C. Markou, & T. Prastitou (Eds.), *EU Internet Law: Regulation and Enforcement*. Cham: Springer International Publishing, pp. 77-98.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S. and Floridi, L. (2016), The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2), doi:10.1177/2053951716679679.

Popper, K. R. (1950), Indeterminism in Quantum Physics and in Classical Physics. Part II. *The British Journal for the Philosophy of Science, 1*(3), pp. 173-195.

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).

Rescher, N. (1958), On prediction and explanation. *The British Journal for the Philosophy of Science, 8*(32), pp. 281-290.

Schellenberg, T. R. (2003), *Modern archives. Principles and techniques*. Chicago: Society of American Archivists.

Schum, D. A. (2001), *The evidential foundations of probabilistic reasoning*: Northwestern University Press.

Selbst, A. D. and Powles, J. (2017), Meaningful information and the right to explanation. *International Data Privacy Law, 7*(4), pp. 233-242. doi:10.1093/idpl/ipx022.

Wachter, S., Mittelstadt, B. and Floridi, L. (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), pp. 76-99. doi:10.1093/idpl/ipx005.

Willis, A. (2005), Corporate governance and management of information and records. *Records Management Journal, 15*(2), pp. 86-97.

World Medical Association. (2013/1964), "Declaration of Helsinki: Ethical Principles for Research Involving Human Subjects", available at: https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/ (accessed 10 September 2019).