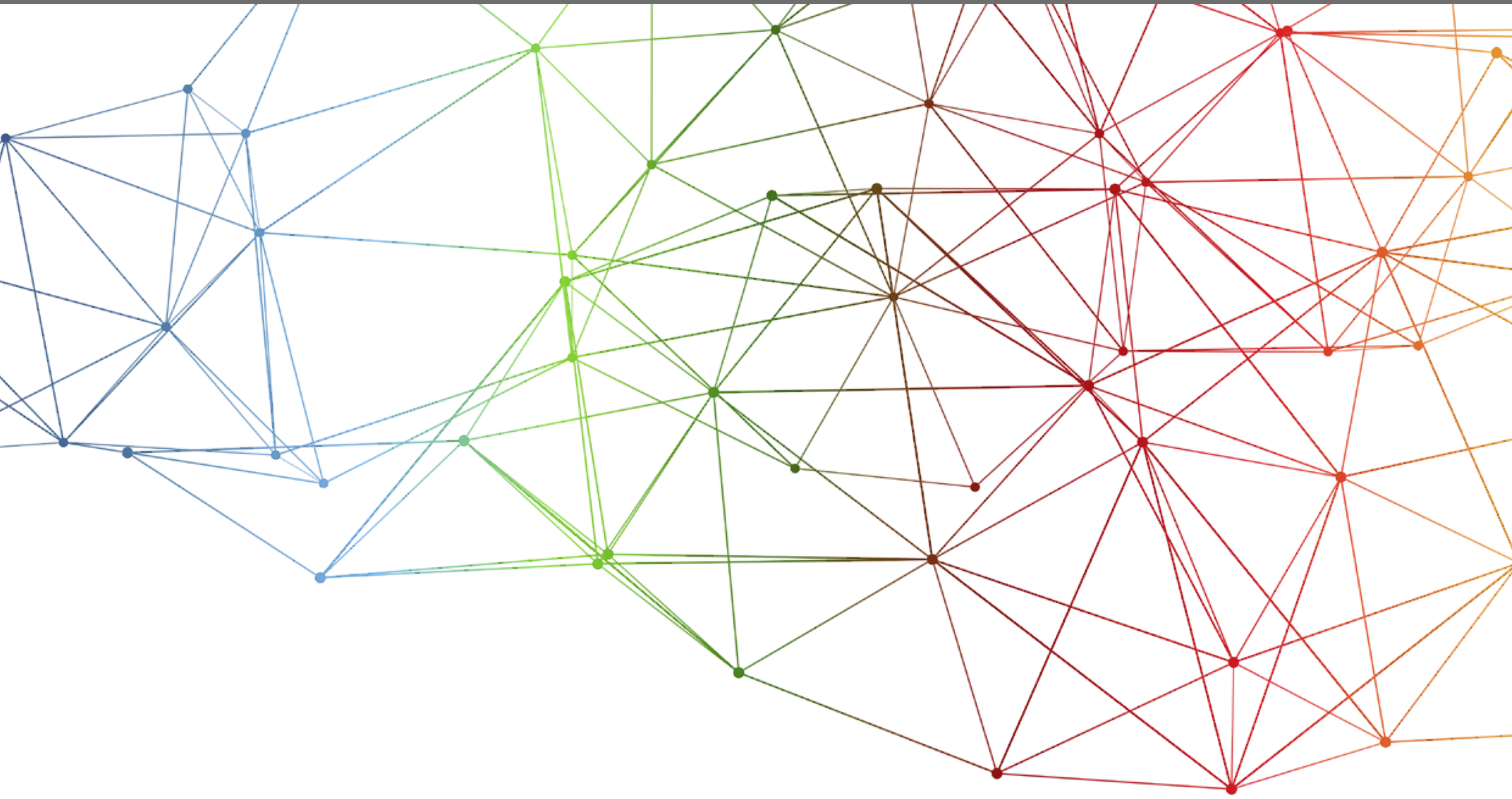
ALGORITHM
WATCH

AlgorithmWatch Working Paper No. 2

Ethics and algorithmic processes for decision making and decision support

by Lorena Jaume-Palasí and Matthias Spielkamp



Executive Summary

Far from being a thing of the future, automated decision-making informed by algorithms (ADM) is already a widespread phenomenon in our contemporary society. It is used in contexts as varied as advanced driver assistance systems, where cars are caused to brake in case of danger, and software packages that decide whether or not a person is eligible for a bank loan. Actions of government are also increasingly supported by ADM systems, whether in "predictive policing" or deciding whether a person should be released from prison. What is more, ADM is only just in its infancy: in just a few years' time, every single person will be affected daily in one way or another by decisions reached using algorithmic processes. Automation is set to play a part in every area of politics and law.

Current ethical debates about the consequences of automation generally focus on the rights of individuals. However, algorithmic processes – the major component of automated systems – exhibit a collective dimension first and foremost. This can only be addressed partially at the level of individual rights. For this reason, existing ethical and legal criteria are not suitable (or, at least, are inadequate) when considering algorithms generally. They lead to a conceptual blurring with regard to issues such as privacy and discrimination, when information that could potentially be misused to discriminate illegitimately is declared private. Our aim in the present article is, first, to bring a measure of clarity to the debate so that such blurring can be avoided in the future. In addition to this, we discuss ethical criteria for technology which, in the form of universal abstract principles, are to be applied to all societal contexts.

Given that the issue of ethics is always also about specific kinds of action and about responsibility for this action, however, it is inevitably dependent on structural and situational contextualization. The rules that apply to the state or to a government, for example, can hardly be applied to citizens as individuals. While this differentiation is standard practice in the realms of ethics and constitutionalism, it has so far been lacking in the debate about automation.

The present paper seeks to contribute to a differentiated ethical and legal debate by introducing a taxonomy. This taxonomy is focused on the issue of action as well as on the dimensions of potential harm inherent in automation. We begin by explaining why technology-neutral ethics are needed. We then look at how actions and decision-making occur within automation before proposing a taxonomy that provides a structure for classifying the various risks and conflicts more appropriately, thus enabling a more differentiated procedure for developing ethical criteria.

We therefore propose a categorization that distinguishes between the category in which algorithmic processes are oriented towards the collective *publicness* (or *social goods*), and we refer to the algorithmic processes dedicated to the individual as the category of *individual goods.* The *social goods* category is divided into the subcategories *societal frame* and *collective goods.*

This taxonomy structures the public/publicness as a dimension in its entirety as a complex structure which cannot be reduced to opinions and information but rather includes collective goods on the one hand and looks at individual and collective interactions on the other. This offers a better ethical contextualization for working out differentiated ethical criteria that highlight a technology-neutral, values-oriented approach.

The development of this working paper was made possible by a fellowship of the Bucerius Lab of ZEIT-Stiftung Ebelin und Gerd Bucerius.

ZEIT-Stiftung Ebelin und Gerd Bucerius

Translation: Kathleen A. Cross

Table of contents

Introduction: Ethics in the digital era

Far from being a thing of the future, automated decision-making managed by algorithms (ADM) is already a widespread phenomenon in our contemporary society. It is used in contexts as varied as advanced driver assistance systems, where cars are caused to brake in case of danger, and software packages that decide whether or not a person is eligible for a bank loan. Actions of government are also increasingly supported by ADM systems, whether in "predictive policing" or deciding whether a person should be released from prison. What is more, ADM is only just in its infancy: in just a few years' time, every single person will be affected daily in one way or another by decisions reached using algorithmic processes. Automation is set to play a part in every area of politics and law.

Current ethical debates about the consequences of automation generally focus on the rights of individuals. However, algorithmic processes – the major component of automated systems – exhibit a collective dimension first and foremost. This can only be addressed partially at the level of individual rights. For this reason, existing ethical and legal criteria are not suitable (or, at least, are inadequate) when considering algorithms generally. They lead to a conceptual blurring with regard to issues such as privacy and discrimination, when information that could potentially be misused to discriminate illegitimately is declared private. Our aim in the present article is, first, to bring a measure of clarity to the debate so that such blurring can be avoided in the future. In addition to this, we discuss ethical criteria for technology which, in the form of universal abstract principles, are to be applied to all societal contexts.

Given that the issue of ethics always focuses on action and responsibility for this action, however, it is inevitably dependent on structural and situational contextualization. The rules that apply to the state or to a government, for example, can hardly be applied to citizens as individuals. While this differentiation is standard practice in the realms of ethics and constitutionalism, it has so far been lacking in the debate about automation.

The present paper seeks to contribute to a differentiated ethical and legal debate by introducing a taxonomy. This taxonomy is focused on the issue of action as well as on the dimensions of potential harm inherent in automation. We begin by explaining why technology-neutral ethics are needed. We then look at how actions and decision-making occur within automation before proposing a taxonomy that provides a structure for classifying the various risks and conflicts more appropriately, thus enabling a more differentiated procedure for developing ethical criteria.

The need for technology-neutral ethics for algorithms

The amount of research dedicated to the issue of ethics in relation to algorithms and other automated processes has increased over the last few years. Some studies (Bozdag, 2013; Naik & Bhide, 2014; Friedman & Nissenbaum, 1996; Tene & Polonetsky, 2013) take a mainly descriptive look at the subjectivity immanent to the programming of algorithms. As these studies demonstrate, machine bias – a set of preconceptions built into a code – is an inevitable outcome of the cultural background and socialization of the developers and data scientists who design and implement algorithmic processes.

As early as 1996, Batya Friedman and Helen Nissenbaum developed a typology of bias in computer systems that described the various ways human bias can be built into machine processes: "Bias can enter a [computer] system either through the explicit and conscious efforts of individuals or institutions, or implicitly and unconsciously, even in spite of the best of intentions". The typology developed by Friedman und Nissenbaum indicates that the way discrimination is standardized by becoming inbuilt within a machine system is not to do solely with the developer or the client. Rather, bias can also arise from contact with the user or from conflicts around the formalization of social phenomena that are hard to formulate in terms of a code. Having observed that bias and discrimination are manifested in machine processes themselves, Friedman and Nissenbaum conclude that ethical analyses of ADM systems should also start at the level of the technology itself.

The basic thrust of this statement is shared by the majority of researchers who have looked at the issue of bias and discrimination in the age of automation. In the following we present three practical variants of the technology oriented approach. In doing so, we highlight what, in our view, are obvious weaknesses of this approach and where it makes extremely presumptive assumptions.

Ethics as a programmable set of instructions for action

The more complex an algorithm is, the more obscure it becomes. In some kinds of machine learning algorithms, the processes developed by the algorithm to generate certain results cannot even be explained by their developers. In such cases, the introduction of first-order (meta-level) algorithms is considered – an ethical authority designed to "supervise" algorithms (Etzioni, Turilli & Wiltshire, 2016; Anderson & Anderson, 2007). This proposal takes rather a lot for granted. It assumes that ethical action can be programmed and automated in a logical language – in other words, that algorithms are capable of thinking, weighing different considerations and, ultimately, performing certain actions. It is also assumed that a kind of ethical programming is possible *without* machine bias. Arguing in a similar vein, Bello and Bringsjord (2012) do concede that moral thinking in algorithms should not be structured along the lines of classical ethical principles because it does not reflect the way in which people make decisions. What they fail to explain, however, is the extent to which algorithms can be said to "think" and why this algorithmic "thinking" should be described as moral.

Ethical criteria oriented toward technology

One and the same algorithm can serve very different purposes. An algorithm used to select a film can be just as useful in cancer research. The way algorithms function must therefore be looked at in context. Both data selection and database as well as the context of application play a role when it comes to the risks and opportunities posed by an algorithmic process; this is especially so in the case of more complex algorithms.

For this reason, other approaches focus on developing normative criteria relating to conceptualization and data processing in algorithmic processes. Many of these criteria come from the realm of data protection (Romei & Ruggieri, 2014; Kamishima 2012). One such criterion is *purpose limitation,* which means that data processing should occur in relation to a clearly formulated purpose and, where personal data is involved, must have a legal basis or the consent of those concerned. Precise purpose limitation along with other principles such as data minimisation, however, mean that certain correlations or criteria in search of patterns are effectively ruled out because they violate data protection regulations.

Transparency – be it in the form of revealing the code (Tutt, 2016) or of being obliged to give an account of how the data is processed algorithmically (Datta et al., 2016; Tene and Polonetsky, 2013a) – is often demanded as a corrective. This demand for transparency is regarded as a *conditio sine qua non* for enabling people to maintain their own "information privacy". Schermer (2011) goes further even than this. While rejecting the concept of data privacy, he calls for a right to group privacy on the part of collectives, as profiling renders the identifiability of individuals irrelevant. More than this, he argues, profiling attempts to group individuals into meaningful categories so that the identity of the individuals themselves is no longer relevant (see also Floridi, 2012; Hildebrandt, 2011; Leese, 2014).

Personal data is not necessarily private data, however. The concept of data privacy reduces the notion of privacy to that of merely having control over one's personal data. In doing so, it overlooks the fact that it is possible to control such data without necessarily enjoying privacy. The essential feature of privacy, however, is the ability to exercise autonomy in relation to political or economic dictates.

> *In the liberal tradition – a key influence on both public debate and the administration of justice in Germany and the European Union – privacy is generally regarded as a condition for enabling autonomy and as its expression, in the sense of being able to think and act independently. (Wehofsits, 2016)*

Having control of one's own personal data is merely one of many ways to achieve autonomy; it is not an end in itself. (Quite apart from which: people had no absolute control over their personal data prior to the digital era, and they still do not have it today. Humans are social animals: whether intentionally or not, they are constantly sharing personal data with those around them.)

When control over personal data is considered an absolute good, however, the idea of normative criteria applied to algorithm design or data processing reveals its flaws: these criteria are considered ethical in nature even though they are merely technical restrictions applied to one or other process – completely detached from any context. They do not refer to human action in the sense of ethics but to the process of conceptualizing and implementing automated data processing. They are not a set of instructions for developers or data scientists but are rather directed at the programme itself. Yet programmes do not act; they perform tasks. Programmes are shaped by the world of their designers and influenced by the individual who tasks the latter. Ethical demands are directed at actors who are in a position to act in the genuine sense of the word.

Ethics of knowledge-related restrictions

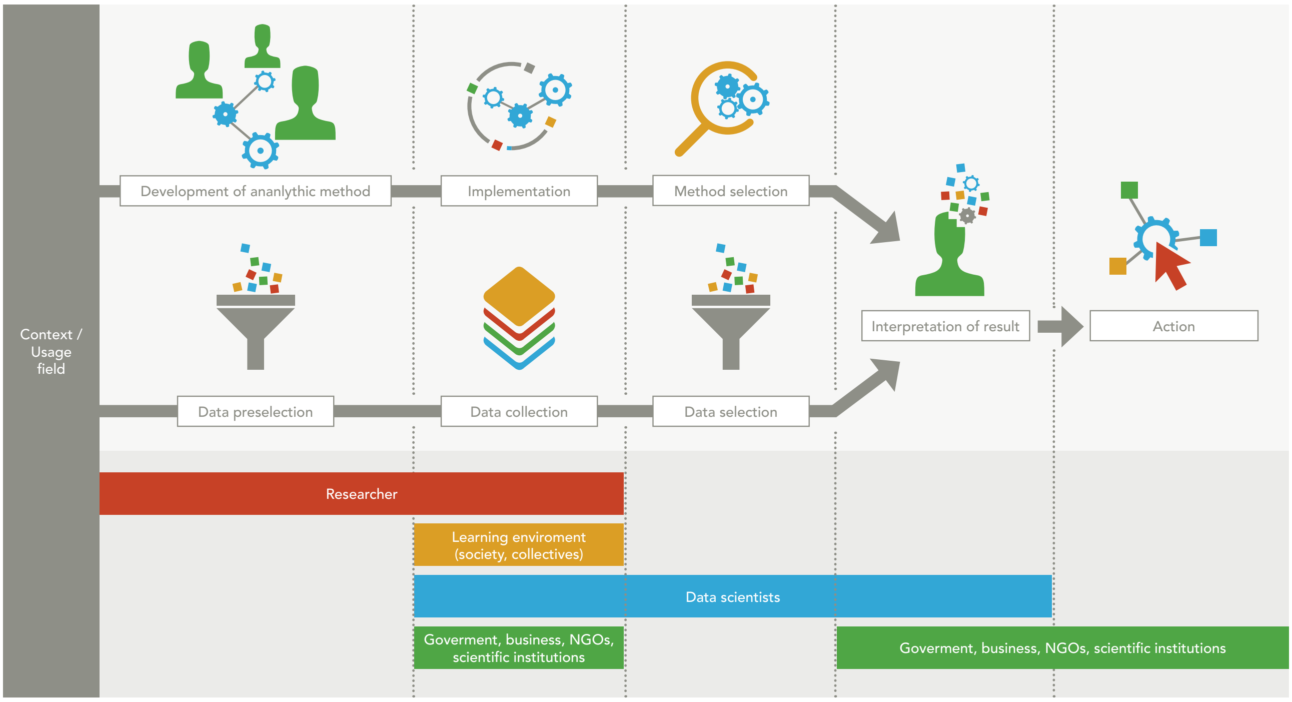
A third variant of the technology oriented approach pursues the strategy of intervening restrictively in the process of knowledge production on which ADM systems are based (Pasquale, 2015). This intervention can occur, for example, by excluding certain data categories from the data processing. Thus, there could be a rule that prohibits the combination of health-related data with financial data in scoring processes, or one that brackets out ethnic heritage when selecting staff or when allocating rental accommodation. Mittelstadt et al. likewise identify risks of discrimination only at the level of knowledge. Here, ethical conflicts are located exclusively in the analytical process undertaken by complex algorithms. These problems may be located in the initial analytical steps, when certain random patterns are erroneously interpreted as significant correlations and thus generate false, incomprehensible or inconclusive evidence. In certain kinds of algorithmic processes, the problems may also be considered to result from a lack of clarity as to the connection between the input data and the correlations resulting from it; this, in turn, may either be because the complexity of the algorithms and the calculations they perform are "black boxed" or because the algorithms have worked with poor data so that the "evidence" derived from them is likewise flawed. Other forms of discrimination occur in relation to conceptualizing the world through algorithms that formalize social relations (such as notions of human dignity) which are difficult per se to capture in a formula (Mittelstadt et al., 2016).

The notion that underlies this approach – that of a "result" of algorithmic processes – is problematic in various respects, however. The results of algorithmic processes (the literature on this focuses primarily on profiling and/or personalization) are patterns identified by means of induction. They are nothing more than statements of probability. The patterns identified do not themselves constitute a conclusive judgment or an intention. All that patterns do is suggest a particular (human) interpretation and the decisions that follow on logically from that interpretation. It therefore seems inappropriate to speak of "machine agency", of machines as subjects capable of bearing "causal responsibility" (Floridi, 2012). While it is true that preliminary automated decisions can be made by means of algorithmic processes (regarding the ranking of postings that appear on a person's Facebook timeline, for example), these decisions are the result of a combination of the intentions of the various actors who (co-) design the algorithmic processes involved: the designer of the personalization algorithm, the data scientist who trains the algorithm with specific data only and continues to co-design it as it develops further and, not least, the individual toward whom this personalization algorithm is directed and to whom it is adapted. All these actors have an influence on the algorithmic process. Attributing causal responsibility to an automated procedure – even in the case of more complex algorithms – is to fail to appreciate how significant the contextual entanglement is between an algorithm and those who co-shape it.

All these approaches rely on some kind of control or restriction of epistemic factors, in other words, of what can be learned or discovered by means of automation. This is not due principally to considerations of privacy but is rather based on the assumption that certain pieces of knowledge or information could be misused – which, however, effectively pre-empts the options for action that are available to all the actors. Automation and algorithms might, namely, also serve to identify many hidden patterns of illegitimate or undesired inequality. This could be used to identify and manage discrimination. The latter possibility is prevented by the approaches described. In addition, all possibility of ethical action in the real sense of the word is denied. Ethical action does not arise on the basis of a few, limited pieces of information. On the contrary: the more information an individual has at their disposal, the better they can put it in an ethical context and act accordingly in a just way.

Instead of this, certain data with a public component is declared private in order to prevent discrimination. This reveals a tendency to equate personal data (in the sense of data protection law) with private data in order to limit the risks of discrimination. Ethical conflicts – which actually ought to be a subject of societal debate – are thus categorically sidelined.

This will not make the conflicts disappear, however; indeed, it may make them worse. To give a (somewhat pointed) example: a woman's pregnancy is considered to be a personal matter. However, from a particular point onwards a pregnancy can no longer be concealed from public view. Declaring a pregnancy to be not just a per-


Development of ananlythic method
Implementation
Method selection
Context / Usage field
Data preselection
Data collection
Data selection
Interpretation of result
Action
Researcher
Learning enviroment (society, collectives)
Data scientists
Goverment, business, NGOs, scientific institutions
Goverment, business, NGOs, scientific institutions

A variety of actors are involved in the process that leads from conceptualizing through developing to applying automated decision-making systems. All these actors have an influence on the results and thus bear part of the responsibility for them.

sonal but a “purely private” matter (with the aim of preventing potential discrimination against pregnant women) would make it necessary to conceal pregnant women from public view and to exclude them from all social participation for the duration of the pregnancy. What is more, such a policy would make women no longer a “normal” sight in public places. There would no longer be any wider social debate about pregnancy and inclusion in public life (for example in the workplace). One could imagine similar scenarios in relation to homosexual people, to people of colour or people with a visible illness. In all these cases, conflicts would not be resolved by privatizing or suppressing information. It would only make it impossible to identify and keep a watch on discriminatory behaviour.

To draw some interim conclusions: all three approaches discussed thus far rely on restrictions enacted via technological processes. This applies equally to so-called supervisor algorithms, to approaches relating to data processing (such as purpose limitation) and those relating to data collection (such as making it illegal to correlate certain data categories with one another). This is presumably done on the assumption that focusing efforts on technological processes offers sufficient abstraction and contextualization to enable ethical principles and restrictions to be formulated for the long term. What this assumption fails to take into account, however, is that the very foundations of digital technologies themselves are constantly subject to change – whereas social conflicts have, in terms of their basic structures, barely changed throughout the course of human history. Ethical principles that are oriented towards technology cannot be ethical per se. They are focused merely on technical processes aimed at preventing unethical action. Human action does not figure in these processes at all. Yet it is exactly this – (good) human action – that is the object of ethics in the true sense of the word.

In the following section, we show that this is not merely a matter of conceptual hair splitting or a question of definitions; rather, algorithms actually do lack fundamental capabilities at the technical level which are necessary for responsible action.

Can decisions and ethics be encoded?

Before considering what are appropriate ethical categories for automation processes, we need to define the terms *decision* and *responsibility*. In a second step, we need to examine whether or not ethical criteria can be encoded in algorithmic processes.

Human decision-making and responsibility

Human action (or indeed any kind of action) is rooted in intention. Action is regarded as behaviours that are controlled and motivated by intentions (intentionality).

When considering the issue of intentionality, it is necessary in turn to distinguish between decisions and motivations (Nida-Rümelin, 2013; Anscombe, 1963; Bratman, 1987 und 1991). Motivations arise from expectations regarding certain consequences and not from a certain kind of action. Decisions, by contrast, are considered as implemented once the relevant actions have been taken, regardless of whether or not the expected consequences have been fulfilled.

Actions are behaviours in which the person concerned bears responsibility for controlling their intentions. A person controls their intentions when they are based on reasons. Ultimately, a person only has control over their actions when they can give reasons for them. For attributing responsibility, a person's reasons, convictions and actions need to be looked at together and tested for coherence. Coherence between convictions, reasons and actions is a structural form of rationality. It is not a matter of the rational content of the convictions but of the procedural rationality of the action. Whether or not a person bears responsibility, then, depends on the degree of their rationality – as well as on the degree of their freedom to act. Thus, responsibility is a "gradual" concept: the person who engages in an action can bear more or less responsibility depending on their state (of mind and body) and on the context in which they have made a decision. A person's emotional state, their state of health, their age (underage vs. of legal age), and the specific options available to them (how much freedom and how many alternatives does the person have?) are relevant factors, ethically and legally, in attributing more or less responsibility. In this sense, a person acting autonomously is never an absolutely autonomous being but rather exists in a certain relation to the matter at hand and to the wider societal context; as such, this person is – at least according to external perceptions and ethical standards – dependent on these factors.

An individual's freedom, then, based as it is on their (rational, emotional) intentions, needs to be underpinned by (good, or many) reasons (Saake, Nassehi 2004). Making decisions accordingly is an expression of that person's free will. And yet no matter how much trouble an individual goes to in order to justify their decision, their free will is not absolute. Free will is not pre-social. It is contextual and – as Hegel noted – reconciles necessity with insight. According to Armin Nassehi (2011) time too is a constitutive element of this will.

> *Free will is socially formed will. [...] We should not will everything; rather we should will that which is in accord with our own cultural image of human nature. Algorithms can help shape the social will and can change it permanently. Algorithms would have us believe that this notion is independent of all context. While we can assume that human nature (the anthropological) is fixed, inseparable from time, will is evidently, factually – and ethically – in flux nonetheless (Nassehi, 2011, 260)*

When considering the responsibility of machines or algorithms and reflecting on ethical criteria for algorithms, then, we need to examine the issue of machines' rationality and freedom (to act).

Rationality and freedom (to act) in the context of automation

Algorithms are mechanisms that lead to certain kinds of results. The processes that occur within more complex algorithms are causal in nature. Action, however, is based on reasons and on the freedom to choose from various options the one that corresponds to the desired decisions or to certain motivations. Responsibility for actions is always attributed in relation to these reasons and this freedom. Reasoning is a logical process. It is not possible, however, to formalize reasonably complex logical processes using algorithmic processes. This has been acknowledged as an incontrovertible fact in Philosophy generally and in the Philosophy of Science in particular ever since studies by Church and Turing emerged in the 1930s. Algorithmic processes cannot prove, for example, the truth of a formula – one of the most elementary components of logic – such as “Socrates is a person”. “Socrates is a person” is a first-order predicate logic formula. First-order predicate logic is a branch of mathematical logic. Predicate logic is concerned with formalizing arguments and testing their validity. It is extremely important in disciplines such as information science, mathematics, linguistics and philosophy. First-order predicate logic is concerned more specifically with the logical inferences arising from certain mathematical expressions. This inferring occurs at a purely syntactical level, i.e. it bears no relation to mathematical significance. First-order predicate logic has led to a number of important insights in mathematics as a whole as well as in philosophy.

Predicate logic formulas can be true in some worlds and false in others. In the 1930s Church and, shortly after him, Turing both proved that algorithmic processes such as the Turing machine cannot test first-order predicate logic theorems in terms of their validity. Church and Turing have not been refuted to date. The Turing machine can prove neither that “Socrates is a person” is true (as the sentence is not universally valid) nor that “Socrates is not a person” is true (as this sentence likewise is not universally valid). If logic is regarded as an important component of reasoning, then reasoning is not an algorithmic process.

Causal relations are, to be sure, algorithmic, but reasoning is not a causal process; it is a logical process in which not only the causes count (i.e. their formal existence) but the substance of reasons, i.e. the actual arguments put forward.

> *If […] we accept a certain understanding of causal relations, which claims that causal relations are algorithmic (i.e. that if I have exact knowledge of the current state of affairs and all relevant laws, I can determine the next state of affairs), then it has been proven since the 1930s that reasoning, in which logical inferences play a role that amount to the complexity of first-order predicate logic, is not a causal process (Nida-Rümelin, 2014)*

If machines or algorithmic processes are unable to perform complex logical operations, they certainly will not be able to take account of ethical issues either. They lack the kind of rationality that is crucial for attributing responsibility. Furthermore, they lack freedom. Algorithmic processes are not capable of making decisions autonomously. Only beings that are free to act are autonomous, self-determined agents. This freedom is expressed in moral actions – in behaviour that can be justified and (to paraphrase Aristotle) originates in the agent themselves, regardless of third parties or external conditions – and, accordingly, is independent of immoral passions, as these are relational in nature (envy, anger, etc.). It follows from this, then, that mechanical processes themselves cannot be attributed responsibility.

Thus, responsibility can only be attributed to those involved in directing and designing algorithmic processes (see table p. 5). However, even responsibility is conditioned by the graduality of what we can know – by the available options for action and by what we can control. More complex algorithmic processes in which those involved cannot foresee how the process will develop and can only control and shape it to a certain extent are currently regarded as potential risk situations and are preventively normed, often by means of absolute prohibition. In the discipline of decision theory, the situations described above are considered to be situations of uncertainty. This way of looking at things enables a different way of dealing with them: in decision theory, if uncertainty exists it should be dealt with according to the principle of minimizing the greatest imminent harm or even ensuring it does not occur at all.

In this respect, responsibility can also be attributed to those individuals who have an opportunity to influence or control a specific algorithmic step – even if they are not responsible overall for the design and development of the process (such as the team of algorithmists and data scientists who work at Facebook). Opportunities for influencing and controlling algorithms are sufficient to justify a responsibility to intervene, much like the obligation to intervene in the sense of having a duty of care (such as the duty to provide assistance in case of an accident).

Ethics and digital geography: A taxonomy

nI terms of regulation, processes in which data is processed in an automated way are initially classified as personal and non-personal processes; they are then divided into subcategories according to contexts of application (health, finances, transport, etc.). If these processes make use of personal data or data that can be traced to individuals, they are regulated primarily by data protection laws.

Algorithmic processes that use databases containing no personal data can be just as important to society, however. They are capable of steering human collectives either directly or indirectly. This therefore requires a different kind of categorization – one that structures more appropriately the various ways people are directly affected by algorithmic processes and that is oriented towards the social context of application rather than the technology-based level of data processing.

In addition, when devising ethical criteria relating to processes of automation, no fundamental distinction is made between criteria for collectives and those for individuals. Ethical criteria for collectives follow a different kind of legitimation than those relating to situations involving individuals. For example, when considering the issue of personal property from an ethical standpoint, it needs to be treated differently than the issue of property belonging to a community, where individuals' claims may be withdrawn in favour of the collective. Furthermore, as the example of the commons shows, individuals cannot resolve ethical problems at the collective level using ethical criteria that focus on the behaviour of individuals. A further aspect is that, as far as individuals are concerned, ethical criteria are often defined in relation to basic rights (such as privacy). It is a different matter altogether with collective goods and framework, where there are no such comparable rights to which ethical criteria might be applied. The need for protecting *publicness* therefore requires a different framework for reasoning.

We therefore propose a categorization at the first level that takes account of this distinction. We call the category in which algorithmic processes are oriented towards the collective *publicness* (or *social goods*). And we refer to the algorithmic processes dedicated to the individual as the category of *individual goods.*

With regard to this categorization it might be countered that a large majority of algorithms that are used by individuals are personalization algorithms. They are not interested in specific individuals per se but rather in patterns that group the individuals into different kinds of profiles. It would therefore be perfectly reasonable to include these too in the category of publicness – and thereby to call into question the meaningfulness of the categorization itself. However, we are talking about two fundamentally different logics of action here.

Theories such as “we-rationality” (Smerilli, 2008) render plausible the notion that individual decisions taken for the good of a collective need not be driven by hidden self-interest, with regard to either the reasoning or the psychological motivation behind them, but rather may genuinely be directed toward the common good. According to the theory of “we-rationality” there are two modes of rationality according to which people act: “we-rationality”, oriented towards the common good and the collective, and “I-rationality”, focused on one's own personal interests. Each mode of rationality can rule the other one out. The theory explains why individuals who act against their own interests are not acting irrationally but are taking into account a different level of interests and intentions in their actions. Thus, individuals make use of different rationalities for individual and for collective interests. An ethics directed toward collective goods and collectives follows a different kind of logic than that of individualistic interests. As in the case of predictive policing (see below), individualistic ethics is not sufficient to even out the societal inequities of the example. It therefore also appears promising to consider decision theories in the context of we-rationality when it comes to working out ethical criteria that take account of the collectivist aspect of algorithmic mechanisms.

The ethical (and legal) debate to date – critically summarized in the above overall review of the literature – focuses on situations concerning individuals' rights. In contrast to this, concerns relating to the common good are rarely incorporated into ethical debates about algorithms and artificial intelligence. And this despite the fact that algorithmic processes are influenced primarily by a collectivist approach. As with the issue of discrimination, ethical conflicts in algorithmic processes are collective in nature: discrimination happens at the level of the individual but is not directed at a specific person. Assignment to a collective is both reason and principle here. For example, Mr. M. is not harassed by neo-Nazis as Mr. M. but because of the darker colour of his skin and the fact that he is assigned to the collective known as "refugees". Paradoxically, the social construction of collectives in such cases is both the trigger for discrimination as well as the referential parameter for checking patterns of discrimination. Modernity responds to discrimination by conceding individual rights to all citizens. Individual rights are not enough, however, to provide structural protection to certain collectives. Indeed, many of the problems that arise as a form of collective discrimination cannot be addressed by reference to individual rights. Viewed in this perspective, predictive policing that uses algorithms which do not process personal data can throw an entire city into a state of social imbalance when these algorithms contribute de facto to creating so-called no-go areas: a disproportionate police presence can suggest a massive security problem rather than more security. In such cases, individuals are neither being discriminated against (they can move somewhere else within the city) nor are they affected in terms of data protection law. In other words, ethics and laws that fail to focus on collectives as groups and their logics have blind spots that conceal a large proportion of the problems and risks associated with automated procedures. For this reason, it seems necessary to extend the categories used to date in order to assign facts and responsibility in more appropriate ways in algorithmic processes.

Publicness (social goods)

Publicness is always related to collectives. The presence of different societal frames that determine both the extent and the forms of interaction of a collective is indispensable for the emergence of the latter. These societal frames can generate both inclusion and division in a society. Individuals are able to exercise their basic rights within these frames. Societal frames also form a point of access to collective goods.

The category of publicness thus includes algorithmic processes concerned with interactions of collectives (by means of which a societal frame is established) as well as with collective goods. We therefore divide this category into the subcategories

- societal frame and
- collective goods

These subcategories are necessary because it is possible to identify ethically relevant differences between the two regarding the attribution of responsibility.

a) The societal frame

The societal frame is the "gateway" to public interaction, to exercising certain basic individual rights (such as freedom of opinion and of association) and for access to collective goods. The societal frame can take various forms: it is the main square in a village where people go shopping for their vegetables on a Saturday, where children play and friends meet one another. It can also be a social platform such as Facebook or Twitter, where parties are organized, fan pages set up, or groups of Twitter users are aggregated in lists. A search engine such as Google – where information is arranged, partly according to personalized relevance, in a rank order – is also a societal frame. Platforms use algorithmic processes that regulate and thereby shape these interactions.

The societal frame is a kind of infrastructure that facilitates and regulates a society's access to publicness and collective goods. A town with no pavements (found widely in the US) offers a different kind of access than a city such as Amsterdam with its pavements, bicycle paths and roads. Transportation networks with their rules and communication networks are also included in this concept. Digital platforms also constitute a kind of societal frame: they are the new digital marketplace. Their formats and algorithmic processes facilitate communication and render collective goods such as knowledge accessible. (Incidentally: the way in which state security is structured and enacted is also part of the societal frame.)

The societal frame has two main characteristics, which make it necessary to distinguish between collective goods and societal frame:

- **Control and capacity to shape:** The task of controlling and shaping the societal frame is in the hands of a limited group of individuals. This applies both to a city or town's architecture as well as to transport and communications networks, to search and timeline algorithms, and to security. These tasks are not shaped and controlled collectively. Rather, they are not available to this open kind of shaping and influencing. This is especially the case with those societal frames (city architecture, security) which are provided by the public sector authorities to facilitate participation and access for everyone. They are subject to strict rules and duties of accountability. This raises the question of whether societal frames not provided by the public sector should likewise be subject to special rights and duties.

- **Access:** The societal frame determines which parts of a collective receive what kind of access to a.o. collective goods and in what way individual rights are exercised in the sphere of publicness.

b) Collective goods

Collective goods are the second subcategory of publicness captured by algorithmic processes.

Collective goods, rather like the terms "common good" and "justice", cannot be defined in terms of their substance. They are too dependent on the context of a given society to do so.

From a normative point of view collective goods must be accessible to everyone, available for use by everyone and capable of being shaped (designed, ordered) by the collective. The factors described here differ in essence from the factors described above in relation to the societal frame. This gives rise to far-reaching consequences regarding allocation of duties and attribution of responsibility. Whereas the societal frame is the responsibility of a limited, clearly identifiable number of actors, the attribution of responsibility in the case of collective goods is more complicated (such as in the case of algorithmic open source processes that are developed further by a broad community). Moreover, the individuals of the collective not only have the right to use the collective good: this right goes hand in hand with a duty to contribute to its continued existence.

Collective goods also differ from societal frames with regard to whether or not they are replaceable. Certain collective goods, such as knowledge, are not replaceable. Any piece of information that is deleted cannot be replaced by alternative information. In contrast to this, a societal frame can become obsolete and be replaced by a new kind of access to publicness. Search engines or social media platforms that are currently successful, for example, may well be replaced at some point by other formats. At the same time, the various societal frames have a certain monopoly – or, to put it another way, competitors within their format are not as prominent. Google, for instance, occupies a certain monopoly position within its format, as do Facebook and Twitter in theirs, and all three offer different modes of access and forms of interaction. And they are just one dimension of publicness, one form of access to it: analogous modes of access to publicness offer a range of services that cannot be equated with digital services.

To decline to interact with publicness is to decline to participate fully in society.

Individual goods

We subdivide individual goods as follows:

- self-selected services (such as fitness trackers, games or music apps)
- services that relate to individuals but are used by third parties (e.g. scoring or support systems used to make a decision on granting a visa)

Individual goods have a different contextual societal frame. With regard to a large proportion of algorithmic goods or services self-selected by individuals, it is generally possible to do without them or to use an alternative. The possibility of declining to use these services, however, gives no indication of their level of sensitivity. Individuals have a relative degree of control over the services. Thus, conflicts that arise in relation to them require weighing the rights of two subjects against one another.

In contrast to this, algorithmic services that affect an individual but are used by third parties are rarely services from which the individual concerned can withdraw. These kinds of services may be used by public sector authorities (e.g. to support decisions on approving social welfare benefits or to support an embassy in granting a visa) but may also be used by private sector actors (credit scoring). In some cases, the individual is not even aware that these services exist. This means that, unlike self-selected services, the individuals affected have hardly any control over these kinds of algorithmic services.

Summary: Ethics and its structural context

The current ethical debate and its associated legal demands are based on modern ideas of democracy and individual rights. Yet algorithms demonstrate that, for example, forms of discrimination can emerge that do not affect the rights of the individuals, as the above-mentioned example of predictive policing shows; rather, discrimination only becomes visible when comparisons are made between different collectives. The opportunities and uncertainties that accompany automation are not solely to do with the issue of discrimination, however. A range of other issues also arise which are to do with all kinds of fundamental rights (to freedom, to equality, to participation). However, algorithms contain a fundamental element that is collective in nature, and this is not being taken into account currently at all, either in ethical or in legal terms, even though it exerts an influence on all these basic rights. The use of logic based on individual rights means in some cases that ethical gaps are overlooked while in others situations are wrongly understood, leading to a false or at least uncertain attribution of responsibility.

The taxonomy outlined here is intended as a structure for developing ethical principles for algorithmic processes – principles which take account not only of the individual character of automation processes but also of their collective character, and which integrate different logics (common good versus individual rights). The aim of this is to enable a more complex and more complete range of power asymmetries and risks of misuse to come into focus. Are we dealing, for example, with conflicts and risks that affect the common good or collectives? Or are they more to do with the public frames in which collectives interact and which provide access to collective goods? Do they create an illegitimate power gap between the state and its citizens? Or is it a matter of weighing the interests of two non-state actors against one another? Which tasks do these actors perform? Are there tasks related to the common good that cannot be automated? Are these tasks relevant to society?

Questions like these also constitute the foundations of the legal categories found in states based on the rule of law. Traditionally, societies governed by the rule of law distinguish between civil law and public law. For constitutional and democratic reasons, the ethical criteria that apply to the state-citizen relationship cannot be applied to the relationship between private individuals or entities – whether with regard to citizen-to-citizen relationships in a narrower sense or to relationships between citizens and companies which, as organizations of private individuals, are likewise bearers of fundamental rights (even if certain kinds of private companies are (or should be) subject to special rights and duties). Interactions between the state and its citizens function, in regulatory terms, according to different legal dogmatics than the interactions between citizens. Thus, for example, the basic rights of the citizen are directly valid as a form of defensive right against the state, as a means of constraining the latter's power over its citizens. In the realm of civil law, by contrast, where laws regulate interactions between citizens, these basic rights are valid only indirectly, as the basic rights of citizens regularly clash with one another and therefore need to be weighed against one another – assuming that the state is called upon at all to act sovereignly in favour of a basic right within the societal frame of its duty of protection.

The classification behind our categories covers two aspects. First, it is necessary to distinguish fundamentally between the

a) ethical logics and
b) structures

of society as a whole (publicness) and the

a) individual logics and
b) structures

that affect individual goods.

Second, we seek to offer a taxonomy that is compatible with above legal distinctions and that can be used to assess the need for regulation and, where necessary, to supplement this regulation. The division of individual goods into self-selected services and services used by third parties, for example, is already present in various legal corpora (e-Commerce Directive, data protection regulations etc.): self-selected services are largely regulated by the legal principle of consent, whereas services that are used by third parties are regulated by instruments such as the monitoring of terms and conditions (in the case of non-state actors) or by specific legal principles (in the case of both state and non-state actors).

At the same time, the category of publicness opens up a new dimension in some respects. Agents that provide – and thus effectively regulate – the frame in which publicness occurs possess a special measure of power and responsibility. So far, this frame has been regulated primarily by state actors, meaning both institutions and actors such as the police or border control agencies as well as, say, legislation regarding the right to demonstrate and to express an opinion, and so forth. In their role of contributing towards the formation of public opinion and providing information, the press and broadcasters have been some of the few non-state actors to date that have similarly played a major role in forming the overall realm of publicness as well as special sub-realms of the same, and have been capable of bestowing or withdrawing visibility in relation to societal affairs. However, this function has been structured by the statutory legal framework: due to their special role, the press and broadcasters have been granted special privileges but have also been made subject to special obligations. Indeed, for a long time, broadcasting was regulated purely by statutory means in the sense of a state-mediated pluralism. The advent of digital media means that new actors are joining the realm of publicness which are taking on neither the role of the press nor any state functions.

That the category of publicness is a necessary one is demonstrated by the fact that there is as yet no consideration of further aspects of the relevant context, in particular of the various ways in which harm or disadvantages may accrue to a society. Such harm may not always affect individuals, as in the case of predictive policing. They may have either an individual impact on collectives or a structural influence on society as a whole by virtue of the fact that they include the subcategories *collective goods* or the *frame of publicness* (societal frame).

In addition, the regulation of public and private space is only partially placed under review in the course of contextualization – the more so given that digitalization and ideas of public and private are fundamentally decoupled from ideas of space. For a long time the legal and ethical debate was – and to some extent, in an extrapolated form through case law, still is – based on the theory of spheres which regards the private and the public as fundamentally separable spaces. Since the idea of space is voided by the internet, debate has focused on ways of dividing these so-called spheres. What is overlooked by this is that the criterion of space used to distinguish the private from the public is a one of a formal nature, and in no way reflects societal ideas of the private and public in their complexity. Due to this focus, a closer look at the structures that make up publicness has been neglected. Issues regarding the provision of access and the structuring of publicness have been regulated only in specific realms (transport, the press, etc.), without looking fundamentally at publicness per se. And issues regarding the societal collective and its function within publicness have only partially been addressed, such as in debates about religion or about the right to demonstrate.

These are the very issues, however, that have come to play a key role as a result of automation processes and that need to be looked at from an ethical and, in some respects, from a legal point of view. The taxonomy presented here takes this as its point of departure, structuring publicness in categories that do not refer simply to opinions or information but include collective goods on the one hand while looking at individual and collective interactions on the other, as well as considering publicness in terms of its structures, points of access and moderating role. In this way, a better ethical contextualization is achieved for the purpose of elaborating differentiated ethical criteria. Such a contextualization serves to place an approach centre-stage that is technology-neutral and oriented towards values.

Bibliography

Anderson, M., Anderson, S.L. (2007). Machine ethics: Creating an ethical intelligent agent. *AI Magazine,* 28(4), 15.

Anscombe, G. E. M. (1963). *Intention,* second edition, Oxford: Blackwell.

Arendt, H. (2002). *Vita activa oder Vom tätigen Leben,* third edition, Munich: Piper, 1967.

Bratman, M. (1987). *Intention, Plans, and Practical Reason*, Cambridge, MA: Harvard University Press.

Bratman, M. (1991). *Cognitivism about Practical Reason, reprinted in Faces of Intention,* Cambridge: Cambridge University Press, 1999.

Bello, P., Bringsjord, S. (2012). *On how to build a moral machine.* Topoi, 32(2), 251–266.

Bozdag, E. (2013). Bias in algorithmic filtering and personalization. *Ethics and Information Technology,* 15(3), 209–227.

Etzioni, A., Etzioni, O. AI Assisted Ethics (2016). *Ethics and Information Technology,* 18(2), 149-156. Retrieved from https://ssrn.com/abstract=2781702 (accessed 2 October 2016).

Floridi, L. (2012). Big data and their epistemological challenge. *Philosophy & Technology,* 25(4), 435–437.

Friedman, B., Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems (TOIS),* 14(3), 330–347.

Hildebrandt, M. (2011). Who needs stories if you can get the data? ISPs in the era of big number crunching. *Philosophy & Technology,* 24(4), 371–390.

Kant, I. (1977). *Werke in zwölf Bänden.* Band 7, Frankfurt am Main.

Leese, M. (2014). The new profiling: Algorithms, black boxes, and the failure of anti-discriminatory safeguards in the European Union. *Security Dialogue,* 45(5), 494–511.

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

Naik, G., Bhide, S. (2014). Will the future of knowledge work automation transform personalized medicine? *Applied & Translational Genomics, Inaugural Issue,* 3(3), 50–53.

Nassehi, A. (2011). *Gesellschaft der Gegenwarten: Studien zur Theorie der modernen Gesellschaft II.* Berlin: Suhrkamp Verlag.

Nida-Rümelin, J. (2014). Agency, technology, and responsibility. *Politica & Società,* 2, 185-200.

Nissenbaum, H. (2001). How computer systems embody values. *IEEE Computer,* 34, 118-120.

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms that Control Money and Information.* Cambridge: Harvard University Press.

Rawls, J. (1979). *Eine Theorie der Gerechtigkeit.* Frankfurt: Suhrkamp Verlag.

Romei, A., Ruggieri, S. (2014). A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review,* 29(5), 582–638.

Saake, I., Nassehi, A. (2004). Die Kulturalisierung der Ethik. Eine zeitdiagnostische Anwendung des Luhmannschen Kulturbegriffs. In: Günter Burkart and Gunter Runkel (eds.): *Niklas Luhmann und die Kulturtheorie,* Frankfurt/M.: Suhrkamp: 102-135.

Schermer, B.W. (2011). The limits of privacy in automated profiling and data mining. *Computer Law & Security Review,* 27(1), 45–52.

Smerilli, A. (2008). We-thinking and ‚double-crossing‘: frames, reasoning and equilibria, MPRA Paper, University Library of Munich. Retrieved from https://mpra.ub.uni-muenchen.de/11545/2/MPRA_paper_11545.pdf (accessed 2 October 2016).

Tene, O., Polonetsky, J. (2013). Big data for all: Privacy and user control in the age of analytics. Retrieved from: http://heinonlinebackup.com/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/nwteintp11&section=20 (accessed 2 October 2016).

Turilli, M. (2007). Ethical protocols design. *Ethics and Information Technology,* 9(1), 49–62.

Tutt, A. (2016). An FDA for algorithms. SSRN Scholarly Paper, Rochester, NY: Social Science Research Network. Retrieved from: http://papers.ssrn.com/abstract=2747994 (accessed 03 November 2016).

Wehofsits, A. (2016). Die ethische Dimension von Big Data. Vodafone Institute. Retrieved from http://www.vodafone-institut.de/wp-content/uploads/2016/10/Big-Data_Ethische-Fragen.pdf (accessed 7 November 2016).

Wiltshire, T. J. (2015). A prospective framework for the design of ideal artificial moral agents: Insights from the science of heroism in humans. *Minds and Machines,* 25(1), 57–71.

Authors

Lorena Jaume-Palasí is co-founder of AlgorithmWatch and focuses her research on the philosophy of law and politics in the digital era. Among other fields, she concentrates on the contemporary idea, dynamics and ethics of digital publicness and privacy. As an external expert she gave testimony to Google's Advisory Council on the so-called Right to be Forgotten and is a founder of the Dynamic Coalition on Publicness of the United Nations Internet Governance Forum. Currently, Lorena is a Bucerius Fellow of ZEIT-Stiftung. She also serves as the head of the secretariat of the German Internet Governance Forum (IGF-D) and on the expert advisory board of the Code Red initiative against mass surveillance. Lorena is co-author and editor of several books on Internet governance and digital policy.

Matthias Spielkamp is co-founder of AlgorithmWatch and a founding member and now publisher of the online magazine iRights.info – about legal issues in the digital world, which in 2006 received the Grimme Online Award, Germany's most prestigious award for online journalism. Matthias gave expert testimony to committees of the German Bundestag on artificial intelligence and robotics, government surveillance, future developments of journalism, and copyright regulation. Currently, Matthias is a Bucerius Fellow of ZEIT-Stiftung; in 2015/16 he was a Fellow at Stiftung Mercator and associate researcher at the Alexander von Humboldt Institute for Internet and Society. He serves on the board of the German section of Reporters Without Borders and is a member of the advisory councils of the Whistleblower Network and the Politics & Public Affairs program at Quadriga University. In the steering committee of the German Internet Governance Forum (IGF-D), he acts as co-chair for the academia and civil society stakeholder groups. Matthias co-authored and edited several books on Internet governance, journalism and copyright regulation and holds master's degrees in Journalism from the University of Colorado at Boulder and Philosophy from the Free University of Berlin.

Legal notice

AW AlgorithmWatch gGmbH
Bergstr. 22
D-10115 Berlin
Registered Office: Oranienstr. 19a, D-10999 Berlin
Register Court: District court Berlin (Charlottenburg)
HRB 186522 B
Executive directors: Lorena Jaume-Palasí, Matthias Spielkamp
email: ljp@algorithmwatch.org | ms@algorithmwatch.org

License

When referencing this publication, please use the following citation:
Lorena Jaume-Palasí, Matthias Spielkamp: Ethics and algorithmic processes for decision making and decision support, AlgorithmWatch Working Paper No. 2, Berlin 2017

The development of this working paper was made possible by a fellowship of the Bucerius Lab of ZEIT-Stiftung Ebelin und Gerd Bucerius.

ZEIT-Stiftung
Ebelin und Gerd
Bucerius

Translation: Kathleen A. Cross

ALGORITHM WATCH

The more technology develops, the more complex it becomes. AlgorithmWatch believes that complexity must not mean incomprehensibility (see our ADM manifesto). AlgorithmWatch is a non-profit initiative to evaluate and shed light on algorithmic decision making processes that have a social relevance, meaning they are used either to predict or prescribe human action or to make decisions automatically.

HOW DO WE WORK?

Watch

AlgorithmWatch analyses the effects of algorithmic decision making processes on human behaviour and points out ethical conflicts.

Explain

AlgorithmWatch explains the characteristics and effects of complex algorithmic decision making processes to a general public.

Network

AlgorithmWatch is a platform linking experts from different cultures and disciplines focused on the study of algorithmic decision making processes and their social impact.

Engage

In order to maximise the benefits of algorithmic decision making processes for society, AlgorithmWatch assists in developing ideas and strategies to achieve intelligibility of these processes – with a mix of technologies, regulation, and suitable oversight institutions.

https://algorithmwatch.org