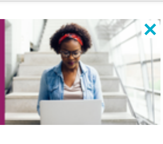
WILEY Online Library

Advertisement
WILEY
Are you a researcher in Germany?
Did you know you could publish your research open access at **no charge** to you?
Find out more


Journal of Law and Society / Volume 50, Issue 1 / p. 100-119
ORIGINAL ARTICLE | Open Access

Human–algorithm hybrids as (quasi-)organizations? On the accountability of digital collective actors

ANNA BECKERS, GUNTHER TEUBNER

First published: 10 March 2023
https://doi.org/10.1111/jols.12412

About | Sections

Abstract

Models of individual accountability for algorithms' actions fail when a human–algorithm association comes to be viewed as a collective actor. In some situations, human and algorithmic actions are so closely intertwined that there is no longer a linear connection between the emergent collectivity and the complex interactions of humans and algorithms. In such collective decision-making sequences, individual accountability can no longer be attributed. Therefore, a new perspective on human–algorithm associations that captures their emergent properties and organizational qualities is needed to develop appropriate models of collective accountability. This article seeks to answer a number of questions. How can the encounter between humans and algorithms within such a socio-technical configuration be adequately theorized? Can the configuration itself be understood as a hybrid collectivity? Can actions be attributed to the configuration as a personified collective actor? How will accountability be institutionalized for human–algorithm associations – in centralized or distributed collective forms?

1 PROBLEM CONSTELLATIONS

In certain situations, communications between humans and algorithms become so closely intertwined that simple systems of interaction are transformed into (quasi-)organizations.[1] A relevant case is 'algorithmic journalism'. Here, algorithms and human actors are brought together in closely timed iterative workflows.[2] As a consequence, algorithmic and human contributions to the jointly authored text are often so closely interwoven that it becomes impossible to identify a responsible author. A strange hybrid emerges, a human–algorithm association.[3]

There are other cases of such hybrids. Spectacular constellations include those of 'digitized corporate governance' – that is, the assignment of management tasks to autonomous algorithms.[4] For example, the company Deep Knowledge Ventures appointed an algorithm as a board member whose task was to communicate with the other members via predictions and other data.[5] Within companies, algorithms can become directly integrated into the collective decision making of the organization.[6] Project organizations or functional units are set up for this purpose, often also in the form of separate organizations. Some companies even use algorithms as an integral part of senior management.[7] Sometimes, they serve as independent board members within a corporate structure;[8] sometimes, they form independent algorithmic sub-organizations, such as subsidiaries.[9] The integration of algorithms in decentralized autonomous organizations (DAOs) goes even further.[10] Here, algorithms independently take over the organization, administration, and decision making of investor groups. In these cases, algorithms do not merely assist in decision making, but themselves act as autonomous decision makers.

Beyond these novel developments, a classic case of close human–algorithm interaction is the cyborg that is characterized by closely interlocking algorithmic impulses and human decisions.[11] However, the media-theoretical interpretation of cyborgs as 'extensions of man'[12] is inappropriate because it conceives of information and participation exclusively from the viewpoint of the human subject, so that the algorithm appears only as an annex of human action capacities.[13] Yet, this is only one out of several possibilities. In some cases, algorithmic calculations clearly dominate human decisions, but in others, it may be the other way around. Furthermore, from a sociological perspective, the interaction between humans and algorithms is never an expansion of the human action capacity; instead, it is a new kind of human–algorithm collective behaviour that emerges.[14]

Empirical studies have shown that organizations entirely controlled and managed by algorithms are today still only hypothetical. Even in the two most extreme cases of autonomy, Plantoid and Metronome,[15] a small amount of human action is still needed. Figure 1 illustrates the present state of management algorithmization.

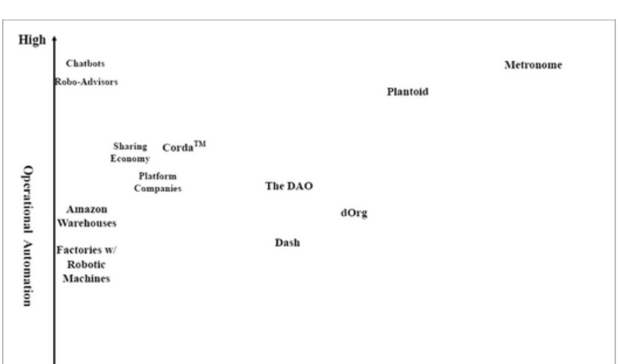


FIGURE 1 Open in figure viewer | PowerPoint

Autonomous business map.

Source: C. L. Reyes, 'Autonomous Business Reality' (2021) 21 *Nevada Law J*. 437, at 470

Hence, despite much speculation in the literature about full algorithmization, it is mostly human–algorithm associations that are identifiable. In numerous instances, algorithms are used as mere tools (bottom left of the figure), and in some rare constellations, almost full algorithmization (top right) can be observed. In our context, cooperation between autonomous algorithms and humans is of special interest (middle area). Here, the mass occurrence of hybrid associations stands in contrast to widespread notions of autonomous algorithms acting in isolation.[16] The close interconnection of the individual actions of humans and algorithms makes it impossible to attribute action exclusively to the human actors involved. Nor is it convincing in these cases to attribute accountability for errors only to the individual algorithms themselves, as proponents of an 'electronic person' ('e-person') do.[17] However, how, then, can we understand human–algorithm associations and which concept of accountability is adequate for them?

To answer this question, it is necessary to break down the broad concept of accountability, which commonly has manifold and overlapping meanings, into two components: responsibility (*Verantwortung*) in the sociological sense and liability (*Verantwortlichkeit*) in the legal and moral sense.[18] From a sociological perspective, responsibility refers to 'the absorption of uncertainty as a social process: that someone draws conclusions from information, communicates them to another, and the next person is then no longer guided by the information but by the conclusions'.[19] When algorithms make decisions, uncertainty absorption no longer takes place in communication among humans, but in human–machine communication. To say it rather bluntly, though inaccurately, the algorithms take over responsibility. This, however, can only happen when they have decision-making autonomy and capacity to act under conditions of uncertainty.

Responsibility is not to be equated with liability. The latter is rather the attempt to reformulate problems of responsibility – that is, problems of uncertainty absorption – in the language of normativity, above all of law, but also of morality, politics, and everyday language. This normative language allows for correcting errors in assuming responsibility. Responsibility for uncertainty absorption is 'translated' into the legality/illegality of a decision and of rights, duties, reproachability, and sanctions. This necessarily creates a divergence between responsibility and liability.[20] This divergence is reinforced when algorithms are used, because law's 'conceptual readiness' is not prepared for the 'opportunity structures' of decision making by algorithms.[21] Law needs to react to the algorithmization of responsibility by changing the rules of liability.

Yet, the attribution of responsibility and liability is even more complicated because human–algorithm associations lead not to only one but to two different emergence phenomena. Emergence is understood here not in the usual meaning as emergence 'from below' – that is, that the interaction of existing system components gives rise to unforeseeable new properties.[22] Instead, emergence occurs 'from above' in the sense of systems theory.[23] A new autopoietic system emerges if, drawing on existing materials, it constitutes its own elementary operations, which in turn constitute the new system in a circular way.

In our context, two questions thus arise. The first question is whether, under certain circumstances, the encounter of humans and algorithms gives rise to an interaction system in the strict sense.[24] This occurs only if the interaction system reconfigures the calculations of algorithms into genuine communications as its elementary operations. The second question is whether, in the transition from a human–algorithm interaction to a human–algorithm association, an organization emerges as a novel autopoietic system that reconstitutes the communications of humans and algorithms differently – namely, as collectively binding decisions within the association as its elementary operations. In both constellations of emergence, accountability will probably be attributed differently.

This article approaches the problem of accountability in human–algorithm associations through four sub-questions that build on each other:

1. Does communication (in the strict sense of systems theory) take place between human actors and autonomous algorithms within human–algorithm encounters, so that an interaction system emerges? Can a personification of algorithms be identified in this context?
2. If so, do such human–algorithm interactions condense into hybrid (quasi-)organizations and under what conditions, and does this change the algorithms' personification?
3. If so, will the ability to communicate be attributed not only to the individual algorithms, but also to such hybrids themselves – comparable to formal organizations?
4. If so, what kind of collective accountability will be institutionalized in human–algorithm associations?

2 THE PERSONIFICATION OF ALGORITHMS

2.1 Algorithms as 'actants'

In formal organizations, two forms of communication attribution are commonly practised. Some communications are attributed to members personally as their own actions. They bind only the members themselves as persons and establish their individual accountability. Other communications, however, are attributed to the organization itself as a collective actor under certain conditions. As collectively binding decisions within the framework of membership, they function as elementary events of the organization. They bind all communications in the entire system, provided that certain infrastructural prerequisites (such as decision-making procedures and rules of representation) are given. This is the basis for the ability to act as a collectivity, which provides the organization with positional gains in environmental relations, because now the organization itself, and no longer only its members, can communicate with its environment. This ability simultaneously establishes collective accountability of the organization as an independent social system.[25]

However, as soon as algorithms are integrated into organizations and are expected to make decisions on their own responsibility, the question arises as to how this affects individual and collective attribution of liability. To this end, the attribution of action to individual algorithms needs first to be clarified. This question cannot be avoided by seeing algorithms only as automated tools of the humans involved. If algorithms display a certain degree of autonomy, decision making, and learning capacity, it is insufficient to understand them simply as an annex to human capacities for action. Their contributions can no longer be identified with the actions of the humans involved. Instead, the question arises as to whether the actions should be attributed to the algorithms themselves who are now identified as 'persons'.[26]

However, the premise that algorithms are addressed as persons is regularly criticized for equating computers with humans.[27] In order not to fall for a false anthropomorphism, one needs first to understand the personification of single algorithms. Here, we can observe a parallel to the action capacity of other non-human actors. In particular, the constitution of formal organizations as legal persons is informative here. We must move away from the familiar idea that legal persons' social substrate is a multiplicity of real people. Nor do the usual suspects come into consideration as collective actors – neither von Gierke's 'real association personality',[28] nor Durkheim's 'collective consciousness',[29] nor Coleman's 'resource pool',[30] nor Hauriou's 'institutions'.[31] Instead – talk incorporated! The collective actor, as defined by Luhmann and other systems theorists, is not a group of individuals but a chain of communications.[32]

The social reality of a collective actor, the social substrate of the legal person,[33] arises under two conditions. First, such a chain of communication must communicate about itself and establish its identity through this self-description. Second, the communicative events produced in the context of this identity must be socially attributed as acts performed by it. When these conditions are met, the social reality of a collective actor emerges. As a semantic artefact, it is the 'person' that turns into the point of attribution for actions, resources, and responsibilities. This rules out reducing collective action to the actions of individuals, as methodological individualism would prescribe.

In a parallel manner, the personification of algorithms is now to be understood. Software agents, robots, and other digital actors are mathematically formalized flows of information. The objects of personification are algorithms that interact with their environment. They are not merely static mathematical formulae but dynamic processes that follow programmes in their problem-solving operations. After an in-depth discussion of various definitions of an algorithm, Gurevich arrives at the following, which emphasizes the processual character:

> A sequential-time algorithm is a state transition system that starts in an initial state and transits from one state to the next until, if ever, it halts or breaks … In particular, a sequential-time interactive algorithm … is a state transition system where a state transition may be accompanied by sending and receiving messages.[34]

Today, social identity and agency are already attributed to such algorithmic processes in the economy and society under certain conditions.[35] However, whether the conditions of personification are fulfilled is not decided in a technology-deterministic way by the choice of a digital technology, but only in the emerging 'socio-digital institution' in whose framework the algorithms are used. Socio-digital institutions are stabilized complexes of social expectations, especially expectations about newly emerging risks, which are formed when digital technologies are used in different social situations. Such institutions are not identical to either social systems or formal organizations. Rather, social systems, including formal organizations and inter-individual interactions, produce expectations with the help of their communications. Such expectations, to use a classical formulation, condense into institutions under an *idée directrice*.[36] Expectations are institutionalized when supporting consensus can be assumed.[37] The remarkable stability of institutions in general, however, is only achieved when they build bridges between different systems.[38]

Socio-digital institutions emerge as effective structural couplings between technical and social systems. In these institutions, expectations of an economic, political, legal, and technological nature come together, and it is often difficult to distinguish between the expectations of the systems involved.[39] In a process of co-production of their components,[40] socio-digital institutions integrate expectations about the opportunities and risks of using algorithms,[41] and these institutions are ultimately responsible for whether personhood is ascribed to the processes involved. However, this happens without a predetermined distinction between human and non-human persons and with the possibility of differentiating degrees of personhood.[42]

The exact parallel between formal organizations and algorithms becomes clear when the common misconceptions about the personification of non-human entities are dispelled. We reject the general idea about the personification of organizations as a collection of people transforming into a real association personality.[43] We equally consider incorrect the notion that, in the case of software agents, a computer is transformed into a *homo ex machina*. Instead, one of the most important 'evolutionary early achievements of social systems'[44] is realized for organizations and for algorithms; under certain conditions, institutionalized social practices attribute agency to communication processes.

In actor–network theory, Latour coins the neologism 'actants' for non-humans capable of acting.[45] At the same time, he elaborates the fundamental difference from the agency of humans. However, he defines the actants' agency too broadly as 'resistance'. By contrast, Rossler elaborates a whole variety of meanings associated with the actions of non-human actors and creates a typology of their forms of action.[46] From this typology, this article selects a type of action that focuses much more narrowly than Latour on participation in social communication.[47] As the term 'actants' already entails, there is no anthropomorphizing of digital processes, but, conversely, a de-anthropomorphizing of all-too-human robots and software agents. These remain – as information philosophers emphasize – 'mindless machine[s]'.[48] Nevertheless, through the attribution of action within socio-digital institutions, they become (non-human) members of society.[49] The human–algorithm association is one of these socio-digital institutions that gives the individual algorithms the status of fully fledged organizational members with the associated authorizations and responsibilities.

However, why do some socio-digital institutions confer personhood on algorithms and others attribute to them only the status of tools or treat them as an integral part of the human body? In encounters with non-human entities, especially algorithms, personification is one of the most successful strategies for dealing with uncertainty, especially the unpredictability of their behaviour.[50] Social personification transforms the human–algorithm relationship from a subject–object relationship into an ego–alter relationship. This admittedly does not give ego certainty about alter's behaviour, but their mutual communication makes it possible for ego to choose its behaviour, even if alter itself remains inscrutable:

> The computer or the algorithm is then no longer a technical artefact with attribution potential, but a partner that appears either anthropomorphically in the case of natural language, or as a corporate actor in the case of decision-making algorithms.[51]

Treating the algorithm as if it were an actor thus transforms uncertainty about causal processes into uncertainty about how to understand the algorithm's reaction to the actions of the human. Causal attribution is transformed into action attribution. This is based on the communicative assumption that the algorithm has motives for its actions. The assumption acts as a hard social reality of actor status, insofar as 'human actors regard their non-human co-actors … as equal partners. They ascribe agency to them – and in a symmetrical way.'[52] This enables humans to choose their own actions, interpret the algorithm's reactions, and in turn draw consequences for the next action.[53]

2.2 Communication with digital actants

However, can we really assume that algorithms communicate as autonomous actors, as we expect of real people and organizations? When we cooperate with algorithms in digital hybrids, do we only perceive the behaviour of machines, or do we communicate with them?[54] In other words, is it mere perception or real communication? The usual answer is to look for 'psychological capabilities previously thought to be reserved for complex biological organisms such as humans'.[55] Another answer, however, is provided by animal, robot, and machine ethics after the recent 'relational turn'.[56] According to this, algorithms as such do not possess ontological actor qualities that allow them to communicate with humans. However, as soon as algorithms become an integral part of social relations – as soon as they become members of digital hybrids, as in our case – the attribution of communicative abilities becomes possible.

Sociological systems theory sharpens the focus. For the personification of algorithms in social relations, particular conditions need to be fulfilled.[57] Whether an encounter between algorithms and humans within a socio-digital hybrid leads to the personification of algorithms depends on whether the 'contributions' of the software agents are constituted as communicative events in the strict sense. Based on the linguistic *trias* of locutionary, illocutionary, and perlocutionary components,[58] communication is defined as an operation that combines three aspects: (1) utterance, (2) information, and (3) understanding.[59] It must therefore be possible to identify events that, as 'utterances' of the algorithm, at the same time contain certain 'information' and whose difference is 'understood' in the follow-up communication. Then a genuine social system does indeed emerge in the encounter of people with algorithms. The 'answers' in the form of communications that we receive from software agents to our 'requests' fulfil all conditions required for communication in the strict sense: a synthesis of utterance, information, and understanding.[60]

This would thus be the first phenomenon of emergence mentioned above; the electronic signals of the algorithms form the material on which the emergent social system of human–algorithm interaction draws to generate communications as its elementary operations. It is therefore not a matter of mere acts of perception by which humans read algorithmic operations, but of meaningful communications (in the strict sense) between humans and algorithms.

Of course, these are merely 'asymmetrical' interactions between humans and algorithms. Nevertheless, a genuine self-producing communication system, a human–algorithm interaction, is emerging between them. This is asymmetrical in three ways.

First, algorithms can by no means be granted the mental competences of humans.[61] Their inner workings consist of mathematical operations based on electronic signals. To understand how communication between computers and humans is possible despite their asymmetrical inner worlds, the distinction between 'subface' and 'surface' is very helpful:[62]

> Subface is the technical side of digital media, characterised by the networking and interconnection of causally controlled processes in hardware and software (if/then/other loops). Surface is the likewise technically designed side, but fundamentally open to access by … communication.[63]

For successful communication, only the surface counts; the subface is no longer part of the communication process. The difference between the computers' subface and humans' consciousness, as fundamental as it is, turns out to be irrelevant for our question. Rather, what is crucial is that a communication process evolves outside the highly different inner worlds of humans and computers. At the same time, however, both the digital subface and the human consciousness must be able to sufficiently 'irritate' this communication process.[64] Only then can the synthesis of utterance, information, and understanding be accomplished.

Second, another asymmetry arises because communication in this relationship is not experienced bilaterally. In communication between humans, the double contingency is experienced symmetrically on both sides, because both partners make the choice of their behaviour dependent on the other's choice and presuppose this in their counterpart.[65] By contrast, between humans and algorithms, the double contingency necessary for communication is only experienced unilaterally – namely, only by humans and not by algorithms.[66] However, such a unilaterally experienced double contingency, as we find in the human–machine relationship, does not rule out the possibility of communication. Historically known configurations, such as communication with God in prayer as well as animistic practices in indigenous societies, certainly achieve the synthesis of utterance, information, and understanding.[67] However, they always do so under the condition that a personification of the non-human partner takes place in the event stream, which enables an attribution of action to the 'other'. Personhood arises wherever the behaviour of others is imagined as chosen and can be communicatively influenced by one's own behaviour.[68] Moreover, in our context of human–algorithm associations, the action attribution to the algorithm constitutes it as a person, as a semantic construct, and thus compensates for the second asymmetry in the human–algorithm relationship.

Third, communication is asymmetrical in terms of mutual understanding. If understanding is defined as the ability to reconstruct the self-reference of another in one's own self-reference,[69] then human consciousness can understand the messages of the algorithm. The internal processes of the algorithm, in turn, are likely to lack the ability to understand the human's inner life. However, since such a mutual in-depth understanding of the inner life of humans and algorithms is not at all important for successful communication, this question can be left open. One must clearly distinguish between the process of understanding within the autonomous process of communication on the one hand and the process of understanding in the inner life of the interacting entities on the other.[70] For understanding within the autonomous communication process, it is not crucial whether the algorithm itself understands or not; rather, it only depends on whether the 'answer text' of the algorithm understands the 'question text' of the human.[71] Thus, if the act of communication of the algorithmic member of the digital hybrid comprehends the difference of utterance and information in the act of communication of the human member and reacts to it with its own difference of utterance and information, communicative understanding is completed within the association. To reiterate, this is true without having to determine whether the inner calculations of the algorithm understand the utterance of the human. Here, prayer as communication with God, animistic practices, and communication with animals provide historical evidence of a communicative understanding. It comes about even if the 'other' (probably) does not reconstruct the self-reference of the human's inner life.[72]

To condense what has been said so far into a short formula: algorithms – just like formal organizations – are nothing more than streams of information. They become 'persons' capable of communicating when a social identity is attributed to them in a socio-digital institution, and when, together with the necessary infrastructural arrangements (such as rules of representation), they are effectively attributed with their own responsibility and liability.

3 ACCOUNTABILITY IN HUMAN–ALGORITHM ASSOCIATIONS

3.1 Human–algorithm associations as 'hybrids'

Up to this point, the question has been whether, within a human–algorithm association, genuine communication occurs between its members, and whether the communicative events can be attributed as actions to the algorithmic member. They can, provided that certain conditions are met. In the next step, the question is whether the human–algorithm association itself (and not only a participating algorithm) can be understood as an independent unit of action, as a collective actor, and what this means for the attribution of accountability.

Here, we encounter the second phenomenon of emergence mentioned above. The human–algorithm association, in contrast to mere interaction, represents a new autopoietic system that transforms the communications of algorithm and human into collectively binding decisions as its elementary operations. The elementary operations sought are not merely decisions, for decisions are also made in interactions and in function systems, such as in law or politics. Only collectively binding decisions (within the framework of membership) can be considered elementary events of organizations. This means 'a special kind of relevance of every decision for others – and this in the double sense of binding effect and responsibility'.[73]

There is no need for intermediate stages in the transition from digital interaction to digital association, especially not for 'digital groups'. In terms of systems theory, a group is not to be understood as an independent type of social system, but as a mere 'mode of interaction and accumulation of interaction'.[74] However, what about the participation of autonomous algorithms in loose communities, families, movements, or networks? Indeed, these all involve membership, and in all of them decisions and even decision chains are found. However, the crucial 'constitutional moment' is missing – the collective bond. As soon as arrangements are made for certain decisions to bind the whole system (such as member roles, norms, and decision premises), then the digital interaction system 'transubstantializes' into a digital organization. However, what does 'binding' through algorithmic decisions imply? Algorithms bind the human–algorithm association in the three dimensions of meaning:

1. time binding: algorithms' decisions generate normative – that is, non-learning – expectations about future decisions;
2. substantive binding: their past decisions bind the content of future decisions (similar to judicial precedents in law);
3. social binding: their decisions bind exclusively the member roles within the association (unlike decisions in law or politics).

The question then is whether, in the case of digital hybrids, a double personification is at play. First, we have the personification of an algorithm as an individual member of the association capable of communicating with human members. Second, the association itself is personified as a collective actor. If this is the case, the duplication of agency would have to entail correspondingly complex attributions of accountability.

Again, it depends on what kind of socio-digital institution is established here. If the algorithm is only used to carry out instructions and choose the appropriate means for given goals, and if the contributions of humans and algorithms can be clearly delimited, the institution of 'digital assistance' is relevant. In this case, there is a principal–agent relationship between the human and the algorithm.[75] The algorithm is commissioned as an agent by its human principal and carries out programmed instructions on transactions vis-à-vis third parties on its own responsibility. When analysing principal–agent relationships, economic theories focus on the risks of an agent acting independently and its deviation from the intentions of the principal.[76] By contrast, philosophy and sociology focus on the positive contributions of both partners to enriching the productive potential of the principal–agent relationship.[77] In the institution of digital assistance, agents act autonomously. However, if something goes wrong, the liability for the decisions that they make on their own responsibility is not attributed to them but to the principals.[78]

Such individualistic concepts of accountability fail, however, as soon as the actions of humans and algorithms become so intertwined that there is 'no linear connection between the emergent structures, cultures or behaviours that constitute collectives and the complex interactions of the individuals from which they emerge'.[79] Humans and algorithms seem to evolve into a symbiotic entity greater than the sum of its parts.[80] In this situation, the social embeddedness of algorithms is contradictory to the understanding of isolated 'algorithmic power', and the institution of digital assistance is replaced by a different kind of socio-digital institution: the human–algorithm association. When the individual contributions of humans and algorithms merge in joint decision making, human–algorithm interactions develop novel collective properties.

Are the responsibility and liability of the hybrid itself also established in this socio-digital institution? If so, in what form? Do such close human–algorithm collaborations become digital organizations in which the responsibility for uncertainty absorption by the algorithm is no longer attributed to the algorithm but to the hybrid itself as a collective actor? How is the liability for such hybrid responsibility redistributed? To answer these questions, it is helpful to transform the first tentative proposals for solutions from network theory,[81] artificial intelligence (AI) ethics,[82] AI liability law,[83] and the sociology of responsibility[84] into a coherent concept of accountability for digital hybrids.

Latour's actor–network theory is relevant here for the second time. We already discussed above his concept of actants, which personifies autonomous algorithms and ascribes to them the ability to communicate. Now, we rely on his notion of the 'hybrid', often used only metaphorically, as a theoretical concept.[85] This concept captures the transformation of the cooperation of 'actants' and humans into a 'hybrid' collective actor. The actants' potential to form associations now becomes relevant. There are many situations, Latour argues, in which capacities for action are needed at a higher level than are available to algorithms. Even if actants have the capacity to choose between alternatives, this would still leave them regularly in a position of paralysis. To counteract this, actants are not only equipped with a language and a resistant body, but also with the ability to form associations. The ability of non-humans to act is drastically expanded when hybrids – that is, associations of human actors and non-human actants – are formed. In this situation, the actants become members of hybrids – that is, of fully fledged associations of a new kind. A pooling of resources takes place, as in any other association. The unruliness of the actants is now combined with the communicative abilities of the humans. Distributed intelligence in such social systems compensates for the psycho-systemic competence deficits of non-humans.[86] By combining human and non-human characteristics within hybrids, algorithms can participate (at least indirectly) in political negotiations, economic transactions, and legal contracting.

3.2 Communication with digital hybrids

Systems theory provides additional insights. In particular, it can explain the motives for personifying human–algorithm associations.[87] Economists point to saving transaction costs in multi-party contracts.[88] Sociologists identify coordination advantages in pooling resources.[89] Lawyers tend to emphasize the 'legal immortality' of incorporated objects: church, state, corporation.[90] Systems theorists point to another aspect that is crucial here; collectivizing action drastically changes the environmental relationship of a social system.[91] Social systems, as soon as they are personified and thus acquire the ability to act themselves, gain considerable positional advantages in communicative contact with their environment. It is through formal organization that social systems' capacity to act can be established. If the membership rules authorize a member to act for the whole organization, then its communications bind all members and, at the same time, they count as actions of the collective actor itself.[92] In this way, human–algorithm associations become collectivities that communicate with their environments. As a consequence, personification allows for ascribing responsibility to the collective actor itself.[93]

This changes the quality of the individual communications of the algorithms involved as well as their social position. Their operations are no longer merely communications in interaction, but now belong to the elementary operations of digital organizations. They are now collectively binding decisions that count as precedents for future communications within the system, and the algorithms are now constituted by the collectivity as members. In addition, some of them are constituted as formally entitled 'organs' that represent the human–algorithm association to the external world. Thus, with their personification, human–algorithm associations cross the threshold in the evolution towards higher complexity, something that Luhmann had reserved only for organizations.[94] Not only digital actors, as shown above, but also digital hybrids can now communicate with their environment in the strict sense as independent agents.[95] For the fact that decisions are attributed to them, human–algorithm associations also form institutions of collective action that ensure that certain actions are selected as binding for the association. Hence, digital hybrids create collective bonds that not only internally determine the meaning of individual members' actions but externally bind the hybrid as a whole. In the economy, for example, human–algorithm associations themselves function as market actors (and not merely their algorithmic or human members), to whom acts of payment, ownership, debts, and bankruptcies are collectively attributed and upon whom collective responsibility and financial liability are imposed.[96] Thus, in addition to rules of representation, decision-making powers, and voting procedures, new collective liability would also have to be institutionalized in hybrids. But in what form?

3.3 Hierarchy: centralized collective accountability

Is centralized accountability established here, as it is prevalent in hierarchical organizations? The humans and algorithms involved do act in their own responsibility, but not in their own name, rather as representatives for the hybrid as a new kind of organizational unit that assumes a hierarchically superior position vis-à-vis the members. They do this in the same way as managers in a company. Managers do not act in their own name, but as agents on behalf of their principal – that is, for the organization.[97] Furthermore, conflicts of interest arise between the members and the human–algorithm association, similar to the well-known agency problems in organizations. Agency theory has developed and institutionalized numerous norms to deal with these conflicts of interest,[98] which for the conflicts in human–algorithm associations would need to be established analogously.[99] At first glance, it may seem counterintuitive to attribute human–algorithm associations with their own preferences. However, if autonomous algorithms can choose different means among pre-programmed goals, this potential for conflict inevitably arises.[100] Furthermore, comparable institutional norms – such as duties and liability of management, the *ultra vires* doctrine of representative standing, and the test of representativeness in class actions – are also necessary to resolve the agency problem in digital hybrids.[101]

Personifying the hybrid as an actor in its own right opens up a collective perspective that no longer attributes action exclusively 'personally', either to the individual human or the individual algorithm.[102] In contrast to individual attribution, collective attribution does justice to the emerging properties of the human–algorithm association in a double sense. First, it accounts for the internal dynamics of human–algorithm interaction and the specific effects of their cooperation. Second, it accounts for the new quality of external relations. It is no longer either the individual human or the individual algorithm to which external communication is attributed. Now, the human–algorithm association itself communicates with its environment. In both respects, the novel risks of the inextricable intertwining of human and algorithmic actions are countered by establishing the human–algorithm association as the common point of accountability for actions.

The Arrow theorem also applies to human–algorithm associations.[103] Their collective decisions cannot simply be calculated as an aggregation of the individual preferences of humans and algorithms. The communication process precludes such an individualizing view and creates the risk of intransparency in internal communication that is typical for organizations. Furthermore, the involvement of algorithms in organizations drastically exacerbates the intransparency. Bostrom describes this risk as 'collective intelligence' or even 'collective superintelligence'. Human–algorithm interactions are not fully controllable, leading to 'perverse instantiation', which refers to when an autonomous algorithm, as a member of the hybrid, effectively fulfils the goal set by the human participant, but chooses a means that violates the human's intentions.[104]

Things become even more confusing if not only the distribution of responsibility is involved, but also the attribution of liability to the three centres of action: human, algorithm, and association. As discussed above, in hybrids it is difficult to identify the risk-realizing event and thus the individual bearer of responsibility. In principle, it can no longer be clarified; was the human action faulty or the algorithmic calculation? When the hybrid as such is responsible for uncertainty absorption, new forms of attribution will make it collectively liable.

Moral philosophers discuss the 'collective moral autonomy thesis', describing paradoxical situations in which misconduct can be identified and accountability for it is attributed to a collectivity, though the individual members cannot be accused of any misconduct.[105] Beyond the intentions and actions of the members, the collectivity is supposed to have its own intentions, act autonomously, and develop information about the possible outcomes of its activities. This makes the collectivity the exclusive point of accountability.[106] This paradoxical situation also exists for the human–algorithm association, which develops its own phenomenal inner perspective. It generates its own hierarchy of preferences, social needs, and political interests. All of this cannot be reduced to one of the actors involved.[107]

The crucial question then is whether accountability attribution can cope with the 'cyborg turn' – that is, the development in which persons and things combine to form an unprecedented form of existence that follows the creeping hybridization of humans and non-humans.[108] One possible answer is that society will grant full collective actor status to a human–algorithm association and ascribe to it responsibility for the actions of its algorithmic and human members as well as liability for wrong decisions. This is a serious possibility. It corresponds to the centralized attribution of liability in formal organizations, but has not yet been tested in practice for the special case of algorithmic communication.

Other than a few authors who work on the sociology of organizations, it is mainly representatives of AI ethics who plead for such centralized collective accountability. AI systems, Heinrichs argues, 'will form mixed agents in close cooperation with their human users, systems whose actions are not easily ascribed to the human user or the AI's programming but rather emerge from their ongoing interaction'.[109] Loh and Loh see the human–algorithm association as a hybrid system that becomes self-responsible due to the purpose of cooperation.[110] Attribution theorists reach similar conclusions when they apply Dennett's concept of 'intentional stance' to hybrids as 'a combination of human, electronic and organisational components'.[111]

Indeed, for a future AI law, several authors discuss the possibility of holding digital hybrids liable in a collectivity-centralized manner. For example, Taylor and De Leeuw suggest rethinking personhood from a relational perspective.[112] In their view, hybrid AI systems should be seen as a new form of experience that affects people's emotional, rational, and social relations to each other, to machines, and to the outside world. For this new form of hybridization of human and algorithmic 'acting', they predict new forms of attribution of action and accountability.[113] This has the consequence that hybrids are treated as new kinds of juridical persons that are directly attributed with accountability for the behaviour of the two inseparable actors.[114] It merges individual actions of humans and algorithms into collective actions. In this way, hybrids can create binding contractual ties, and liability claims against them can possibly be recognized.

3.4 Network: distributed collective accountability

Such collective accountability of hybrids, however, requires institutional preconditions that are not (yet) fulfilled today. Economic practice would have to provide hybrids with sufficient financial resources, or the law would have to mandate compulsory insurance. Only then would a centralized form of accountability for hybrids become effective. Therefore, it is more likely that a form of accountability will be institutionalized that is more closely linked to the bilateral–individual relationships between humans and algorithms, but that, at the same time, can account for the multilateral–collective properties. Network theories are relevant here, as they conceptualize digital hybrids as multiple bilateral relationships between humans and algorithms and, simultaneously, as overarching alliances. In addition to the multilateral links, the organizational unity of the hybrid is made accessible to a (limited) personification: 'Networks produce hybrids, and here, non-humans attain aspects of personhood.'[115]

The concept of purpose is the lynchpin that allows network theory to introduce a third form of attribution that differs from both decentralized individual and centralized collective attribution.[116] It conceives of networks neither as markets nor as hierarchies, but as independent social systems that operate between – or beyond – multilateral relations on the one hand and hierarchical organizations on the other. Adopting a concept of purpose could recognize that network participants have to adjust to a paradoxical double orientation.[117] They are expected to make the difficult synthesis of pursuing both their own individual purposes and the overarching network purpose in one and the same operation. This network-specific compulsion to double orientation is in clear contrast to hierarchical organizations, whose management functions are not allowed to be oriented towards individual interests, but exclusively towards the organizational purpose. It also contrasts with purely multilateral agreements, whose purpose of exchange is directed towards the legitimate pursuit of individual interests. Networks are thus characterized by a constant paradoxical oscillation between individual and collective purpose orientation.

Parallel to this, the human–algorithm association could now also be qualified as a decentralized network organization,[118] in which an individual and a collective orientation are simultaneously institutionalized. Machines and humans are each understood as following their own purposes. At the same time, however, they are oriented towards an overarching cooperative purpose that guides the collective human–algorithm actions.

Categorizing human–algorithm associations as networks has implications, above all, for the attribution of accountability. If the aforementioned dual-purpose orientation constitutes a network that includes both algorithms and human participants, a form of collective accountability other than the centralized one is inevitably institutionalized: a genuine network accountability.[119] In AI ethics, Neuhäuser argues in this sense for a collective moral accountability of human–algorithm associations that now become accountability networks.[120]

At this point, we need to return to the distinction between responsibility and liability. Distinguishing between responsibility as uncertainty absorption and liability as normative reaction to errors helps to solve the problem of network accountability. While some action capacities are centralized in the network as a collective actor, the main financial resources and control capacities remain distributed in a decentralized manner among the network nodes. This is why some authors assume that 'social networks are irresponsible entities'.[121] To hold them accountable nonetheless, a two-step procedure is introduced. In the first step, responsibility for uncertainty absorption is attributed to the collectivity. This circumvents the futile identification of individual actions. In the second step, liability is attributed in a decentralized manner. The responsibility for uncertainty absorption is taken up by the association, but due to the lack of its own resources, it cannot be held financially liable. The collective attribution of responsibility to the hybrid thus ultimately serves to channel the attribution of liability to several outside actors.[122] To clarify again, such a decentralization of liability does not mean that the algorithmic or the human actor becomes responsible or liable for their contribution. Rather, the responsibility for uncertainty absorption is collectivized, but the liability for this collective action is to be borne by the nodes of the network.

This leads to the question of who among the network nodes can be held liable for risky algorithmic decisions. The answer of the leading AI philosophers Floridi and Sanders is distributed liability.[123] The effects of AI-based decisions rely on myriad interactions between many actors, including designers, developers, users, software, and hardware. Distributed

many actors, including designers, developers, users, software, and hardware. Distributed agency is followed by distributed liability.[124]

Human–algorithm hybrids are surrounded by a satellite network of independent actors: users, operators, traders, producers, and programmers. If these actors were integrated into a hierarchical organization, it would be collectively liable, as shown above. After all, one of management's tasks is to coordinate the interfaces of various actions. If, on the other hand, the actors were to interact in pure market relationships, the liability risk would shift to the customers, who are responsible for coordinating the partial services themselves. In our case of a digital hybrid with a surrounding actor network, however, attributing liability exclusively to one of the actors involved is not convincing, either to the producer or to the network centre or to one of the network nodes. Such individual attributions of liability are arbitrary, especially if they ignore the structural diffusion of responsibility in cooperative networks.

Consequently, after attributing responsibility for uncertainty absorption to the hybrid, financial liability will be distributed to those network nodes who benefit from the hybrid's activities – that is, operators, owners, manufacturers, and suppliers of the electronic technology. This solution finds some support in AI ethics and AI liability law. Chinen attributes liability for the hybrid's damaging actions to a group of human operators behind the hybrid itself.[125] When software developers, manufacturers, and engineers share the common purpose of producing an autonomous machine, he submits, they can be liable for harms caused by that machine. Allain argues that future legislation should create a new digital liability regime.[126] Restitution will be equally shared among actors to spread the risk of loss better and reduce the economic disincentives. Navas proposes that a conception of (market) share liability could be suitable in case of liability for AI,[127] and Vladeck rightly suggests that such a common enterprise liability would be a form of court-compelled insurance.[128] Similarly, the European Expert Group on Liability for New Technologies recommends joint liability of all actors associated with the 'commercial and technological unit' of the hybrid.[129] Operators, manufacturers, distributors, and programmers of the software agent are bundled in such a liability network. Possibly, a combination of systems and network theory will be able to develop conceptual tools in the future that identify the boundaries of the liability networks as well as the relative participation of the network nodes in the accountability dynamics.

ACKNOWLEDGEMENTS

For criticism and suggestions, we thank Elena Esposito, Veronika Tacke, and the anonymous reviewers.

Download PDF

ABOUT WILEY ONLINE LIBRARY

Privacy Policy
Terms of Use
About Cookies
Manage Cookies
Accessibility
Wiley Research DE&I Statement and Publishing Policies
Developing World Access

HELP & SUPPORT

Contact Us
Training and Support
DMCA & Reporting Piracy

OPPORTUNITIES

Subscription Agents
Advertisers & Corporate Partners

CONNECT WITH WILEY

The Wiley Network
Wiley Press Room

WILEY

Copyright © 1999-2023 John Wiley & Sons, Inc. All rights reserved