The Risk-based Responsibility for Algorithmic Failures

Anna Beckers, Gunther Teubner

Algorithmic failures pose significant risk to society and are capable to create uncertainty and hereby undermine trust in new technologies. Against this background, the EU has started to regulate algorithmic operations by focussing on the risks that they pose. In this contribution, we argue that the EU's risk-based regulation and liability approach is to be welcomed generally but requires adaption regarding the definition of risks and related allocation of responsibility to the various actors involved in algorithmic operations. Rather than focussing on the severity of risk as a benchmark and centre human failures, we propose a risk-based responsibility that focuses on the risks deriving from the integration of algorithms within different socio-digital institutions.

A. Algorithms, Risks, and Regulation: A critique of the European AI Act

Society's increasing reliance on algorithms brings about significant uncertainty. Large responsibility gaps for wrongful decisions appear under current law when autonomous algorithms are employed in decision-making, when algorithms and humans make collective decisions, or when machines operate in an interconnected manner. As a result, people damaged by algorithmic operations have minimal chances of success in obtaining compensation. At the same time, the lack of clearly delineated responsibility subjects challenges the regulation of technology: Who should be subject to regulation? Who should respond to the risks of algorithmic operation?

Furthering trust and mitigating uncertainty via allocating risks have been the main goals of various legislative initiatives, particularly in the EU, in their regulatory approach to new technologies.[1] Such a risk-based approach shifts the perspective: Rather than viewing technology regulation through the lens of specific technical properties or assuming ex-ante legal

1 Marise Cremona, 'Introduction', in Marise Cremona (ed), *New Technologies and EU Law* (OUP 2017) 2.

obligations, risk-based regulation defines as the source of responsibility the tangible social dangers such technologies may create. However, there is still significant uncertainty about how the risk categories should be defined.

In the recently adopted AI Act[2], the EU proposes the severity of the risk as the primary criterion for imposing obligations on actors. The AI Act prohibits systems that carry an unbearable risk, places significant obligations on manufacturers and deployers of so-called high-risk systems, and imposes transparency obligations for those actors involved in other AI systems that do not fall within the two categories. An exception to this risk-orientation is the regulation of general-purpose AI and foundation models. Here, specific technological properties serve as a basis for responsibility. A further differentiation is made according to the type of harm caused by a particular AI system. The literature proposes similar classifications, distinguishing between safety risks and fundamental rights risks.[3]

However, such categorization of risks along the type of damage faces several problems. First, the abstract concept of severity is not sufficiently sensitive to the social context. Generative AI, such as ChatGPT, is a striking example. Whether generative AI produces a high or low risk ultimately depends on its concrete use in a particular context. Generative AI sometimes creates high systemic risks to society; sometimes, its risks are minimal. ChatGPT-produced birthday invitations or out-of-office replies are not particularly risky, while mass production of racist posts on social media creates enormous political damage.[4] The same technology is used in both cases but the risks differ drastically. In addition, classifying risk according to severity may be of little help for the normative allocation of risk to different actors. A classification into high-/low-risk or types of harm does not provide sufficient normative guidance about the person to be held liable should a risk materialize.

2 Regulation 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, OJ L, 2024/1689, 12 July 2024.

3 Cf Christiane Wendehorst, 'Liability for Artificial Intelligence: The Need to Address Both Safety Risks and Fundamental Rights Risks' in S Voeneky, Philipp Kellmeyer, Oliver Mueller and Wolfram Burgard (eds), *The Cambridge Handbook of Responsible Artificial Intelligence: Interdisciplinary Perspectives* (CUP 2022) 189 ff.

4 Philipp Hacker, Andreas Engel and Marco Maurer, 'Regulating ChatGPT and other Large Generative AI Models' (2023) *FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* 1112.

The European approach in the EU, with the AI Act and the complementary liability rules, needs to be criticized even more harshly. While it aims to respond to the highest AI risks, it nevertheless fails to address the truly novel risk that artificial intelligence has brought about – the autonomy of algorithms. Via establishing legal obligations for different risk situations, the EU legislation addresses only human failures in dealing with AI but ignores algorithmic failures that happen independently of human behaviour. The liability rules to which the AI Act refers are mainly fault-based liability[5] and product liability[6]. And here comes the crucial point. When the human actors involved have fulfilled all their obligations but the algorithms make nevertheless wrongful decisions, neither tort liability nor product liability will compensate the victims for the damages.[7] Thus, the ambitious EU legislation fails to remove a large responsibility gap and hereby fail to realise the objective of fostering trust by addressing and mitigating risks.

While a risk approach is to be welcomed in general, the legally relevant qualification of risks should be adapted in two ways. First, risk-based regulation needs to be sensitive to the social context in which technologies are used. And second, it needs to address not only risks stemming from human action of manufacturing, operating, importing or deploying new technologies, but also algorithmic failures. Therefore, we propose a risk typology that addresses both the dangers of autonomous algorithmic decisions and their occurrences in different socio-digital institutions. This typology, we suggest, provides more robust criteria for connecting specific risks with proximate responsible actors and appropriate legal rules. In contrast to

5 Initially, this link between EU-based regulatory duties and national tort liability was explicitly proposed in the Directive on AI Liability, see European Commission, Proposal for a Directive on adapting non-contractual civil liability rules to artificial intelligence, COM(2022) 486 final. With the adoption of the AI Act only and the recent explicit withdrawal of the Proposal for a AI Liability Directive, the exact contours of liability for breach of the AI Act in national liability rules will depend on the interpretation by national courts and the parallel national implementation of the Product Liability Directive. Cf for this interrelation between AI regulation and AI liability Gerhard Wagner, 'Liability Rules for the Digital Age' (2022) *Journal of European Tort Law* 191, 232 ff.

6 See the 2024 revised Directive on liability for defective Products, Directive 2024/2853 of the European Parliament and of the Council of 23 October 2024 on liability for defective products and repealing Council Directive 85/374/EEC, OJ L 2024/2853, 18 November 2024, which now includes the 'ability to continue to learn' as a product defect (Art. 7 (2) (c)).

7 For details see Anna Beckers and Gunther Teubner, *Three Liability Regimes for Artificial Intelligence* (Hart Publishing 2022) 71 ff.

the EU's risk approach, we do not accept that technological properties determine the character and intensity of the risk; instead, risks derive from the technology's concrete application in different social contexts.

At first sight, this may suggest a sector-specific regulation for each concrete type of technology. However, such sector-specific regulation is too fragmented. Common denominators across sectors in the technology's functions need to be addressed. Conversely, an algorithm may be employed for different tasks in the same sector. For example, in the financial sector, we find the delegation of investment decisions to individual robo-advisors and, at the same time, the interconnections of multiple trading algorithms in high frequency trading. Highly diverging algorithmic risks stemming from the different use of technology occur in the same sector. And the other way around, the different forms of usage of AI are not at all sector-specific. Decision-making is delegated to algorithms equally in other contexts and is not specific to the financial sector.

B. Three socio-digital institutions and their risks

We distinguish three different forms of employing algorithms in social contexts and their related social risks – autonomy risk, association risk, and interconnectivity risk. The starting point for our argument is a typology developed in IT studies that distinguishes three types of machine behavior – individual, collective, and hybrid.[8] However, to avoid the technology-deterministic short-circuit of inferring risks, regulation, and liability simply from technological properties, we suggest introducing "socio-digital institutions" as intervening variables between technology and law. Socio-digital institutions mean stabilized complexes of social expectations, which, in our case, are expectations regarding the behaviour of algorithms in social contexts. Such institutions are neither identical with social systems nor with formal organizations, or social relations. Instead, social systems, including formal organizations and interpersonal relations, produce expectations via their communications, which – to use a classical formulation – condense into institutions under an "idée directrice". Such expectations are institutionalized when consensus can be assumed to support them.[9] Now, socio-

8 Iyad Rahwan, Manuel Cebrian, Nick Obradovich and others, 'Machine Behaviour', (2019) *Nature* 477, 481 ff.

9 Niklas Luhmann, *A Sociological Theory of Law* (Routledge 1985) ch.II.4.

digital institutions integrate diverse technical and social expectations about the opportunities and risks of using algorithms in co-production.[10] These institutions serve as effective structural couplings between technical and social systems, including the legal system.

Socio-digital institutions are different from traditional social institutions because of their technicity. Codes and programs now take over the ordering function that previously symbolically meaningful orders would bear.[11] Such new "techno-digital normativity" differs from the normativity generated in human interaction, leading to new risks. A closer analysis of socio-digital institutions provides the criteria for distinguishing between three risk constellations and identifying responsibility subjects that should bear such risks.

(1) The autonomy risk arises from independent "decisions" in individual machine behaviour. It comes up in the emerging socio-digital institution of "digital assistance", which transforms digital processes into "actants". The humanities and the social sciences are needed to analyse how the institution of digital assistance shapes the productive potentialities of the actants and, in particular, the specific risks they pose to principal-agent relations. The "actant" no longer just follows the principal's predefined program but disposes of degrees of freedom that make its decisions unpredictable. The risk consists of the principal's loss of control and exposure to the agent's intransparent digital processes. This raises two questions: Should the law attribute a particular type of legal subjectivity to autonomous algorithms? Which legal rules in contract formation and liability law could mitigate the autonomy risk of digital assistance?

(2) The association risk of "hybrid" machine behaviour arises when activities are inseparably intertwined in the close cooperation between humans and algorithms. In this situation, a new socio-digital institution—"human-algorithm association"—emerges whose sociological analyses will identify emerging properties. Consequently, it is no longer possible to attribute individual accountability to either single algorithms or humans. Instead, legal solutions that account for the aggregate effects of intertwined

10 On the co-production of different social systems Andrew Feenberg, *Technosystem: The Social Life of Reason* (HUP 2017) 75; Sheila Jasanoff, 'The Idiom of Co-Production, in Sheila Jasanoff (ed), *States of Knowledge: The Co-Production of Science and Social Order* (Routledge 2004) 1 ff.

11 Thomas Vesting, *Gentleman, Manager, Homo Digitalis: Der Wandel der Rechtssubjektivität in der Moderne* (Velbrück 2021) 220.

human and digital activities are required, rendering the hybrid association and its stakeholders accountable.

(3) The interconnectivity risk arises when algorithms do not act as isolated units but like swarms in close interconnection with other algorithms, thus creating different collective properties. Here, a new socio-digital institution develops expectations about dealing with society's structural coupling to interconnected "invisible machines". In this case, the distinct risk lies in the total opacity of the interrelations between various algorithms, which cannot be overcome even by sophisticated IT analyses. Sociological theories of de-personalised information flows within such an anonymous swarm of algorithms demonstrate that it is impossible to identify any acting unit, neither individual nor collective. Consequently, the law is forced to give up the identification of liable actors and will need to determine new forms of social responsibilisation.

C. The autonomy risk of digital assistance

I. Socio-Digital Institution: Assistance

We focus on algorithms operating in the "digital assistance" situation. This incipient socio-digital institution determines a specific social status for individual machine behaviour.[12] "Digital assistance" originates in the time-honoured social institution of "human representation." Someone steps in and acts in someone else's place vis-à-vis a third party. This social institution enacts and produces a type of actorship called "representing agency." As opposed to the social role of a messenger, where Alter only carries out quasi-mechanically Ego's strictly defined orders, representing agency gives Alter the general authorisation to make independent decisions in the name of Ego. At the same time, it also determines the limits of this authorisation so that under certain conditions, Alter is barred from speaking and acting for Ego.[13]

Obviously, the transformation of human representation into digital assistance produces new risks. Four more specific risks need to be identified in the general autonomy risk: identification of the agent, lack of understanding between human principal and algorithmic agent, reduction of

12 Rahwan, Cebrian, Obradovich and others (n 8), 481.
13 Katrin Trüstedt, 'Representing Agency' (2020) *Law & Literature* 195, 200.

institutional productivity, and deviation of algorithmic decisions from the principal's intention.

II. Specific Risk: Autonomous algorithmic decision-making

While it is relatively unproblematic in human representation to identify the representing individual, it is frequently difficult to determine the contours of the AI agent that makes the decision in digital agency. Only once an algorithm is carefully shielded from active external input is it clearly identifiable as the agent speaking for its human principal. However, algorithms are rarely totally isolated. Frequently, they rely on external data input for their decisions; thus, they are not entirely detached from the operations of other digital machines. Only when the actual machine behaviour remains linked to the individual algorithm and its use of the data will the institution of digital assistance still govern the participants' roles. The new risk of identification of the 'responsible' algorithm needs to be mitigated not only by evidentiary rules, i.e., to trace back the wrongful decision in a whole chain of calculations, but also by a legal conceptualisation of algorithmic actorship and clear attribution rules. Obviously, this is no longer possible when digital operations are indiscriminately fused with human communications or interconnected with other algorithms to such a degree that no decision centre can be identified anymore. Then digital assistance will be replaced by institutionalised hybridity or interconnectivity. Below, we will discuss these socio-digital institutions and their legal regime.

While in human representation, a mutual understanding between principal and agent in the process of authorisation can be presupposed, this cannot be maintained when humans delegate tasks to machines. Digital assistance as an institution excludes genuine understanding between human minds and algorithmic operations. Instead, understanding is reduced to a one-sided act of putting the algorithm into operation. And even if understanding of mind and calculation cannot happen, understanding is nevertheless possible in concatenating different communicative acts between humans and machines. The advantages of such delegation lie in the abilities of machines to outperform humans in certain types of behaviour, such as handling a large amount of information in a short period. However, the risks of such communicative understanding need to be compensated by a liability regime that shifts action and responsibility attribution from the human to the digital sphere.

The social institution of human representation has a productive potential that is insufficiently understood if representation is described only as mere task delegation from Ego to Alter. Instead, it is the potestas vicaria conferred by the institution of representation that enables Alter to step in and act in Ego's place vis-à-vis a third party.[14] The potestas vicaria is responsible for the productivity of human representation because the agent need not unconditionally follow the principal's intentions. It is not the principal's will that is decisive; it is the project of cooperation between the principal and the agent. This is the very reason why representation constitutes autonomous actorship of the agent.

In the transformation of human representation into digital assistance, there is a risk of losing this productivity potential. The fear of the homo ex machina drives tendencies to narrow down the algorithm's decisional freedom and reduce it to strict conditional programming. However, the institution of digital assistance requires sufficient degrees of freedom for the algorithm so that the relationship between humans and algorithms can develop its creative potential. Blind obedience to the principal will not do. The reduction to the status of sheer tools needs to be ruled out. Not only human but also algorithmic representatives need to be endowed with the "potestas vicaria, in which every act of the vicar is considered to be a manifestation of the will of the one who he represents."[15] The agent acts "as if" he were the principal. Indeed, it amounts to a revolution in social and legal practice when sheer calculations of algorithms bring about the "juridical miracle" of agency law:[16] A simple machine calculation is able to bind a human being as well as create liability for its wrongful actions. The algorithmic agent representing a human being does not only "sub-stitute" but "con-stitute" the principal's actions.[17] One should not underestimate the consequences of such digital potestas vicaria. In comparison to program-

14 Referring to the theological origins of the vicarian relation, Giorgio Agamben, *The Kingdom and the Glory: For a Theological Genealogy of Economy and Government* (SUP 2011) 138 f.

15 Ibid, 138 f. For a detailed interdisciplinary analysis of this *potestas vicaria,* Katrin Trüstedt, *Stellvertretung: Zur Szene der Person* (Konstanz University Press 2022) *passim*, in particular for algorithmic agency, ch V 4.2.

16 See generally: Ernst Rabel, 'Die Stellvertretung in den hellenistischen Rechten und in Rom, in HJ Wolf (ed), *Gesammelte Aufsätze IV* (Mohr Siebeck 1971 [1934]) 491.

17 Menke's thesis that the agent's will con-stitutes and not only sub-stitutes the principal's will makes the dramatic changes involved visible when algorithms are given the power to conclude contracts, Karl-Heinz Menke, *Stellvertretung: Schlüsselbegriff christlichen Lebens und theologische Grundkategorie* (Verlag Johannes 1991).

ming and communicating with computers, digital assistance opens a new channel of human access to the digital world and allows for the use of its creative potential. Here, we find why digital assistance requires the necessary personification of the algorithmic agent and supports technologies that increase degrees of algorithmic autonomy.

But at the same time, digital assistance exposes society to new dangers of non-controllable digital decisions. Notwithstanding the advantages of digital assistance, such representation through the digital sphere is countered by what we call the autonomy risk. The autonomy risk manifests itself when actions are delegated to the uncontrollable digital sphere and thus may lead to damage. Such unpredictability may stem from the particularities of the programmed machine or the data used to train and operate the algorithm. The result is the same: humans do not control the algorithm they have endowed with action capacity. The law eventually needs to respond to this risk of autonomous decision-making by re-orienting its doctrine to fill the liability gaps and deciding on the legal status of such delegation. We will show that the answer is neither equalising electronic agents with humans by awarding full legal personhood nor treating digital assistance as a mere tool. Instead, the answer is to confer limited legal personhood. We conceptualize digital assistance as an agency relationship and thus make an analogy to agency law for algorithmic contract formation. In addition, the rules of vicarious liability become applicable to constellations of digital assistance. These rules respond accurately to digital assistance and the specific roles it creates for humans and algorithms.

Here is the fourth risk of the principal-agent relation, which emerges from an asymmetric distribution of information. The human principal has insufficient information about the algorithmic agent's activities; the algorithmic agent has information unknown to the principal.[18] This opens new insights for the unexpected productivity of digital assistance. The digital agent may devise contractual solutions that the principal had never imagined. While economic theories of principal-agent relations stress the risks of the agent's deviation from the principal's intentions, philosophy and sociology focus on both partners' positive contributions to enriching the principal-agent relation's productive potential.[19] Both aspects need to be carefully balanced in choosing an appropriate legal regime.

18 eg: Dimitrios Linardatos, *Autonome und vernetzte Agenten im Zivilrecht* (Mohr Siebeck 2021) 128 ff.

19 eg: Trüstedt (n 13), 195.

Altogether, the autonomy risk associated with using algorithmic assistants is much higher than the simple automation risk in entirely pre-determined computer systems. The human actors decide only about the computer program and its general use for contract formation, while in numerous single contracts, the software agents make concrete choices effectively outside human control. Even the programmer can no longer determine, control, or predict the agent's choices ex-ante or explain them ex-post. The algorithm's autonomy does not interrupt the causal connection between programmer and contract, but it interrupts the attribution connection effectively.[20]

III. Responsibility attribution: Users/Operators

Digital assistance, which generates responsibilities only within the bilateral relation between the algorithm and the human user/deployer (or organization), needs to be accompanied by a legal regime that assigns the principal, i.e., the user, the responsibility for the wrongfully acting agent. Principal-agent liability does not hold liable the multitude of actors involved in the algorithm's use, i.e., programmers, manufacturers, traders, etc. Instead, it exclusively targets the user who delegates a task to the technology and thus assumes the autonomy risk. Therefore, only the human user/deployer (or organization) is responsible for the algorithmic failures. In contrast, some authors argue that this unfairly shifts all the risks to the user/deployer alone. They also see other actors in the role of the responsible principal, mainly the manufacturer or producer, including the back-end operator who provides program updates and similar services in the background.[21] In doing so, however, they ignore that the user has assumed the specific risk of task delegation. As a result, they arrive at an unfair distribution of risk between manufacturer, programmer, and user. All actors involved in the construction and operation of the algorithm create different types of risks. These risks must be defined precisely in each case and then allocated exclusively to those who have assumed them. Principal-agent liability responds to the dangers of the division of labour between the user and the algorithm.

20 Gerhard Wagner, 'Verantwortlichkeit im Zeichen digitaler Techniken' (2020) *Versicherungsrecht* 717, 724.

21 European Parliament, 'Civil Liability Regime for Artificial Intelligence' *Resolution of 20 October 2020 with Recommendations to the Commission on a Civil Liability Regime for Artificial Intelligence* (2020/2014(INL) P9_TA-PROV (2020)0276), para 8.

In contrast, product liability, which certainly remains applicable, responds to the specific risks of programming, manufacturing, and monitoring the algorithms but leaves considerable gaps in liability.

D. The association risk of digital hybridity

I. Socio-Digital Institution: Digital hybridity

Next to delegating decisions to algorithms, we observe a different relation between algorithms and humans: collective human-machine decisions. Here, attribution of responsibility differs due to the varieties of socio-digital institutions: principal-agent relation versus association. In digital assistance, agents act autonomously. If anything goes wrong, the liability for their decisions is not attributed to them but to the principals. However, such an individualistic concept of accountability fails 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".[22]

A relevant case is "algorithmic journalism". Here, algorithms and human actors are brought together in closely timed iterative workflows.[23] Consequently, 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.[24] There are other cases of such hybrids. Spectacular constellations include "digitized corporate governance" – that is, the assignment of management tasks to autonomous algorithms.[25] For example, Deep Knowledge Ventures appointed an algorithm as a board member whose task was

22 Mark A Chinen, *Law and Autonomous Machines* (Edward Elgar 2019) 101.

23 Konstantin Dörr, Algorithmischer Journalismus – Eine Analyse der automatisierten Textproduktion im Journalismus auf gesellschaftlicher, organisatorischer und professioneller Ebene (University of Zurich Main Library 2017).

24 Nick Diakopoulos, Automating the News (HUP 2019) 15 f.

25 Marcus Becker and Philipp Pordzik, 'Digitalisierte Unternehmensführung' (2020) Zeitschrift für die gesamte Privatrechtswissenschaft 334, 334.

communicating with the other members via predictions and other data.[26] Within companies, algorithms can become directly integrated into the collective decision-making of the organization.[27] Sometimes, they serve as independent board members within a corporate structure;[28] sometimes, they form independent algorithmic sub-organizations, such as subsidiaries.[29] The integration of algorithms in decentralized autonomous organizations (DAOs) goes even further.[30] 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 act as autonomous decision-makers.

Beyond these novel developments, a classic case of close human–algorithm interaction is the cyborg characterized by closely interlocking algorithmic impulses and human decisions.[31] However, the media-theoretical interpretation of cyborgs as "extensions of man"[32] 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.[33] 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 reverse. 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.[34] In such a symbiotic relationship between humans and algorithms, the collective

26 Florian Möslein, 'Robots in the Boardroom: Artificial Intelligence and Corporate Law' in Woodrow Barfield and Ugo Pagallo (eds), Research Handbook on the Law of Artificial Intelligence (Edward Elgar 2017) 649.

27 Hirokazu Shidaro and Nicholas A Christanikis, 'Locally Noisy Autonomous Agents Improve Global Human Coordination in Network Experiments (2017) Nature 370.

28 Möslein (n 26).

29 John Armour and Horst Eidenmüller, 'Self-Driving Corporations?' (2020) Harvard Business Law Review 87, 106 f.

30 Christoph Jentzsch, 'Decentralized Autonomous Organization to Automate Governance', manuscript available at <https://lawofthelevel.lexblogplatformthree.com/wp-content/uploads/sites/187/2017/07/WhitePaper-1.pdf>.

31 Pim Haselager, 'Did I do that? Brain-Computer Interfacing and the Sense of Agency' (2012) Minds & Machines 405.

32 Marshall McLuhan, Understanding Media. The Extensions of Man (Gingko Press 2003).

33 Katharina Block and Sascha Dickel, 'Jenseits der Autonomie: Die De/Problematisierung des Subjekts in Zeiten der Digitalisierung' (2020) Behemoth 109, 111.

34 Rahwan, Cebrian, Obradovich and others (n 8) 483.

association is greater than the sum of its parts.[35] 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.

II. Specific Risk: indeterminable association of human and machine action

The novel collective properties pose novel social risks. The association risk differs from the autonomy risk in relevant aspects. The Arrow theorem, which prescribes that collective decisions cannot be calculated as an aggregation of individual preferences, also applies to digital hybrids. The participation of algorithms intensifies this intransparency. Bostrom analyses this risk under "collective intelligence" or even "collective superintelligence".[36] The human-machine interactions cannot be fully controlled, which leads to "perverse instantiation": an algorithm efficiently satisfies the goal set by the human participant but chooses a means that violates the human's intentions.[37] And the subtle influence of algorithms on human behaviour is even riskier, as the invisibility of the calculating machines as an integral element of the decision-making may conceal where the actual decision has taken place.

When it comes to accountability, the association risk makes it difficult to determine the damage-causing event as well as the responsible individual. Identifying the illegal action may still be possible – errors in journalistic work as defamation, a corporate board decision as breach of fiduciary duties, social media interaction as collective defamation. However, attributing responsibility to an individual contribution is impossible. Was it the human action or the algorithmic calculation that was at fault? The contrast to the autonomy risk we dealt with above is obvious. For autonomous agents' decisions, it remains possible to delineate individual action, violation of duty, damage, and causality between action and damage; here, the algorithm's decisional autonomy creates the liability gap. In digital hybrids, while it

35 Jacob Turner, Robot Rules: Regulating Artificial Intelligence (Springer 2018) 167.
36 Nick Bostrom, Superintelligence: Paths, Dangers, Strategies (OUP 2017) 58 ff, 65 ff,155 ff.
37 Ibid., 146 ff.

remains possible to identify damage and action, the typical responsibility is due to the impossibility of determining the individual actor. The only way out is to consider the hybrid itself a responsible collective actor. And it is this collective decision-making of hybrids that the law needs to respond to.

III. Responsibility attribution: network

In contrast to principal-agent liability, which exclusively burdens the user, digital hybridity allows the wrongful acts to be attributable only to the human-machine association. However, as long as the association does not have its own assets, it is necessary to channel the resulting responsibility to the multitude of actors who are "behind" the digital hybrid. A whole network of different actors is involved in and benefits from the human-machine association. As control in the network is dispersed across the network nodes, liability must also follow this specific risk structure. We consider "network liability" to be well-equipped to assign, in a fair manner, responsibility to the network participants for the digital hybrid's failures.[38]

The digital network liability we propose is modelled on the American "enterprise liability" and the German Gesamthandshaftung. It works in two steps: attribution of action, then attribution of liability. In the first step, the wrongful act is attributed to the hybrid as a collective actor. This avoids the difficulty of identifying the contributions of humans and the algorithms involved. In the second step, liability for the collective action is channelled to the network members. These members have built and controlled the network, even if only indirectly. They profit from its activities. As a result, all network nodes are liable according to their share. The share is determined by economic benefit from and control over the hybrid. In analogy to the well-known market-share liability, we propose a "network-share liability".[39] An exception is only the constellation in which a company centrally coordinates the network based on contractual agreements. Here, primary

38 David Vladeck, 'Machines without Principals: Liability Rules and Artificial Intelligence' (2014) *Washington Law Review* 117, 149; Jessica Allain, 'From Jeopardy! To Jaundice: The Medical Liability Implications of Dr. Watson and Other Artificial Intelligence Systems' (2013) *Louisiana Law Review* 1049, 1074.

39 For a general discussion of network liability, see: Gunther Teubner, *Networks as Connected Contracts* (Hart Publishing 2011) 264 ff, 267 f.

liability should lie with the controlling company.[40] As a rule, this will be the producer, who will then have recourse to the other network nodes.

E. The interconnectivity risk of interdependent digital operations

I. Socio-digital Institution: Exposure to interconnectivity

In contrast to digital assistance and digital hybridity, our third risk situation, collective machine behaviour, is a purely technological matter. It emerges in the interconnectivity of autonomous algorithms without any human interference.[41] Interconnectivity is different from digital assistance because it is impossible to identify an individual algorithm as responsible. It differs from hybrid human-machine associations because society is ultimately exposed to the interconnected algorithms without being able to establish communicative relations. In collective machine behaviour, there is no two-way communication between humans and algorithms, not to speak of an associative relation between them, but only an indirect structural coupling.

The interdependent algorithmic calculations can be qualified as a "restless collective" based on distributed cognition.[42] Such a "collectivity without a collective" cannot be described as a formal organization or a network. It is only a "swarm" of algorithms arising from chance encounters. Systems theory describes society's relationship to algorithmic swarms as social contact with "invisible machines".[43] Their influence on society is difficult to grasp. As said above, there is no genuine communication between humans and algorithms, nor does a communicative collective emerge from humans and algorithms. Instead of a direct influence mediated through communication, interconnected algorithms exert an influence on social relations that is only indirectly mediated through structural coupling. Therefore, applying the le-

40 Rory van Loo, 'The Revival of Respondeat Superior and Evolution of Gatekeeper Liability (2020) *Georgetown Law Journal* 141, 189.

41 If legal analysis identifies a human involvement, the case would qualify as vicarious liability in digital assistance or network liability in digital hybrids. For more details on the three liability regimes, see Beckers and Teubner (n 7) 153ff.

42 Carolin Wiedemann, 'Between Swarm, Network, and Multitude: Anonymous and the Infrastructures of the Common' (2014) *Distinktion: Scandinavian Journal of Social Theory* 309, 313.

43 Niklas Luhmann, *Theory of Society, Volume 1* (SUP 2012) 66; similarly Mireille Hildebrandt, *Smart Technologies and the End(s) of Law* (Edward Elgar 2015) 40.

gal liability rules for individual algorithms or human-machine associations is impossible. Instead, we propose fund solutions that require political and administrative decisions by regulatory authorities, which distribute responsibility to the respective industry.

II. Specific Risk: Interconnectivity

The social risk of interconnectivity lies in the inaccessibility of the calculations and the impossibility of predicting and explaining the results. The authors of the interdisciplinary study on machine behaviour summarise these unexpected properties under the term "collective machine behaviour":

> In contrast to the study of individual machines, the study of collective machine behaviour focuses on the interactive and systemwide behaviours of collections of machine agents. In some cases, the implications of individual machine behaviour may make little sense until the collective level is considered. … Collective assemblages of machines provide new capabilities, such as instant global communication, that can lead to entirely new collective behavioural patterns. Studies in collective machine behaviour examine the properties of assemblages of machines as well as the unexpected properties that can emerge from these complex systems of interactions.[44]

The study group refers to studies on micro-robotic swarms found in systems of biological agents, on the collective behaviour of algorithms in the laboratory and in the wild, on the emergence of novel algorithmic languages between intelligent machines, and dynamic properties of fully autonomous transportation systems. In particular, they discuss huge damages in algorithmic trading in financial markets. The infamous flash crashes are probably due not to the behaviour of one single algorithm but to the collective behaviour of machine trading as a whole, which turned out to be totally different from that of human traders resulting in the probability of a more significant market crisis.[45]

The interconnectivity risk destroys fundamental assumptions constitutive for action and liability attribution. Interconnectivity rules out the

44 Rahwan, Cebrian, Obradovich and others (n 8) 482.
45 Ibid.

identification of individual or collective actors as liable subjects.[46] It does neither allow for foreseeability of the damage nor causation between action and damage.[47] Dafoe speaks of "structural dynamics", in which

> it is hard to fault any individual or group for negligence or malign intent. It is harder to see a single agent whose behaviour we could change to avert the harm or a causally proximate opportunity to intervene. Instead, we see that technology can produce social harms, or fail to realize its benefits, because of a host of structural dynamics. The impacts of technology may be diffuse, uncertain, delayed, and complex to contract over.[48]

Accordingly, legal scholars refer to complexity theory and philosophies of the tragic when attempting to understand interconnectivity and its potential damages.[49] According to complexity theory, linearity of action and causation cannot be assumed, and surprises are to be expected. Unpredictability and uncontrollability result both from sufficient information and from a poorly designed system for which someone can be responsible; they are inherent in complex systems. Latent failures characterise complex systems that are always run as "broken systems".[50] Coeckelbergh compares the catastrophes resulting from interconnectivity to experiences of the tragic.

---

46 See: Herbert Zech 'Liability for AI: Public Policy Considerations' (2021) *ERA Forum* 147, 148 f.; Indra Spiecker 'Zur Zukunft systemischer Digitalisierung: Erste Gedanken zur Haftungs- und Verantwortungszuschreibung bei informationstechnischen Systemen', (2016) *Computer und Recht* 698, 701 ff.; Susanne Beck, 'Dealing with the Diffusion of Legal Responsibility: The Case of Robotics' in Fiorella Battaglia, Nikil Mukerji and Julian Nida-Rümelin (eds) *Rethinking Responsibility in Science and Technology* (Pisa University Press 2014) 167; Luciano Floridi and J.W. Sanders, 'On the Morality of Artificial Agents' in M Anderson and SL Anderson (eds), *Machine Ethics* (CUP 2011) 205 ff.

47 See Curtis EA Karnow ' The Application of Traditional Tort Theory to Embodied Machine Intelligence' in Ryan Call, Michael A Froomkin and Ian Kerr (eds) *Robot Law* (Edward Elgar 2016) 73: 'With autonomous robots that are complex machines, ever more complex as they interact seamless, porously, with the larger environment, linear causation gives way to complex, nonlinear interactions.'.

48 Allan Dafoe 'AI Governance: A Research Agenda', *Centre for the Governance of AI, Future of Humanity Institute, University of Oxford*, < https://cdn.governance.ai/GovAI-Research-Agenda.pdf> 7.

49 Christiane Wendehorst 'Strict Liability for AI and other Emerging Technologies' (2020) *Journal of European Tort Law* 150, 152 f.; Chinen (n 22) 94ff; Karnow (n 47) 74.

50 See generally: Richard I Cook, 'How Complex Systems Fail', Research Paper, < https://how.complexsystems.fail> 4.

Conventional understandings of blame, responsibility, and even causation fall short.[51] Any retrospective identification of a disaster's cause cannot be but "fundamentally wrong", and responsibility attributions are "predicated on naïve notions of system performance".[52]

Many scholars agree that for interconnectivity, neither ex-ante nor ex-post analyses can identify the actors as attribution endpoints and their causal contribution to the damage.[53] European legislative initiatives had been well aware of the difficulties of liability law:

> AI applications are often integrated in complex IoT environments where many different connected devices and services interact. Combining different digital components in a complex ecosystem and the plurality of actors involved can make it difficult to assess where a potential damage originates and which person is liable for it. Due to the complexity of these technologies, it can be very difficult for victims to identify the liable person and prove all necessary conditions for a successful claim, as required under national law. The costs for this expertise may be economically prohibitive and discourage victims from claiming compensation.[54]

Yet, if the attribution of action, causation, and responsibility is impossible, should the law respond to the risks of interconnectivity at all? Once we accept that interconnectivity is inevitably prone to failure, we might conclude that nothing needs to be "fixed" by law. Interconnectivity risks may be a price to pay for the use of technology. However, there is a plausible counter-argument. Despite being invisible, unpredictable in their operations, and incomprehensible in their underlying structure, interconnected systems

51 Mark Coeckelbergh 'Moral Responsibility, Technology, and Experiences of the Tragic: From Kierkegaard to Offshore Engineering', (2012) *Science and Engineering Ethic,* 35, 37. For an application in relation to interconnected autonomous machines: Chinen (n 22) 98f.

52 Cook (n 50), points 5 and 7.

53 Karin Young, *Responsibility and AI: A Study of the Implications of Advanced Digital Technologies (Including AI Systems) for the Concept of Responsibility within a Human Rights Framework* (Council of Europe Study DGI (2019)05, 2019) 62 ff; Klaus Heine and Shu Li, 'What Shall we do with the Drunken Sailor? Product Safety in the Aftermath of 3D Printing, (2019) *European Journal of Risk Regulation* 23, 26 ff.; Herbert Zech, 'Zivilrechtliche Haftung für den Einsatz von Robotern: Zuweisung von Automatisierungs- und Autonomierisiken' in Sabine Gless and Kurt Seelmann (eds) *Intelligente Agenten und das Recht* (Nomos 2016) 170.

54 European Commission, 'Report on the Safety and Liability Implications of Artificial Intelligence, The Internet of Things and Robotics, COM(2020) 64 final, 14.

do produce results that may represent a productive surplus of meaning.[55] They generally result in intended results. Automatic and even more so autonomous infrastructure may be regularly out of control but still fulfils a distinct purpose, which allows for automation of processes, alignment of procedures, and reasonable calculations. This has two consequences: First, digital technology does not require consensual practices of actual people; acceptance originates in its problem-solving capacity. Second, human actors tend to be paralysed when the risks materialise, when complex technological systems do not function, when they go astray and cause damage. This means that society cannot tolerate their malfunctions once it has accepted complex technological systems. Technological risks must be mitigated and their damages compensated, even if no culprit can be identified. Therefore, de-personalised compensatory rules need to counteract the risks of new evolving technologies.

III. Responsibility attribution: Socialising of risk

In the case of interconnectivity, determining who should bear the risk is different—responsibility shifts from those directly involved to a larger social collective. The interconnectivity of "invisible machines" makes it impossible from the outset to determine an individually responsible algorithm. Since there is only an indirect "structural coupling" between algorithmic interconnectivity and society, no one-to-one responsibility relationship can be established. Therefore, we propose that liability funds be established. The funds should be financed by the industry sector involved.[56] The players' contributions are calculated based on their market share and specific problem-solving capacity. The US Superfund for environmental damage can serve as a model here.[57] The Superfund aims not only to compensate individual affected parties but also to provide rules for remedying the broader social and ecological impact, including regulations on clean-up and prevention. This idea should be taken up for algorithmic interconnectivity. Restitution measures will serve as additional instruments of liability law. In the case of large-scale damage, the regulatory authority responsible

55 See: Armin Nassehi, *Patterns: Theory oft he Digital Society* (Polity Press 2024), 141 ff.

56 Olivia J Erdélyi and Gabor Erdélyi, 'The AI Liability Puzzle and a Fund-Based Work-Around, (2021) *Journal of Artificial Intelligence Research* 1309.

57 42 US Code § 9601 ff.

for the fund should be empowered to select actors with a robust problem-solving capacity and impose the task of restitution and undoing adverse consequences. The actors involved are obliged to take measures that limit or even eliminate the negative externalities of interconnectivity for the future, such as reversibility[58], creation of firewalls or slowing down of interconnectivity or, ultimately, the shut-down of dangerous technological systems, described as the "death penalty" for robots.[59]

F. Conclusion

With these three categories of socio-digital institutions, risks, and responsible actors, we thus shift the focus for risk specification for AI regulation. In contrast to the European Union's AI Act and the related technology rules, which define risks primarily based on damage severity or technical properties criteria and human/organisational obligations, we suggest defining risks according to the social context in which the technology is used.

The AI Act distinguishes between the obligations of various actors in the "AI value chain"[60] but, without consistent explanation, links such actor obligation to the risk severity of technology or considers, as in general-purpose AI and foundation models, the specific technological properties because of their general riskiness as a reason for imposing obligations. It differentiates between the respective responsibilities of providers, importers, distributors, and deployers/users but does not justify why a particular actor is supposed to bear the risk.

Instead, we suggest that risk and responsibility should not be defined according to damage severity but according to the institutional context of the technology's application. Users – or deployers in the AI Act – have a specific responsibility when they delegate decision-making to algorithms. Funds should be created to manage the systemic risk of interconnected algorithms. The network of actors creating AI is mainly the risk-bearing

58 See on reversibility European Parliament, 'Resolution of 16 February 2017 with Recommendations to the Commission on Civil Law Rules on Robotics', P8_TA ("017), OJ 18/7/2018, C252/239, Annex.

59 Mark A. Lesley and Brian Casey, 'Remedies for Robots' (2019) *University of Chicago Law Review* 1311, 1390.

60 On the term of the AI value chain and the problem of responsibility attribution Jennifer Cobbe, Michael Veale and Jatinder Singh, 'Understanding Accountability in Algorithmic Supply Chains' (2023) *FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* 1186.

collective when human and algorithmic decision-making aggregates into collective decision-making.

To end with some examples: We do not propose a specific regulatory framework for general-purpose AI. Instead of an ex-ante allocation of responsibility among manufactures, providers, and deployers/users, the responsibility would be allocated with a view to the specific context. The famous case of a lawyer letting ChatGPT write court briefs is a delegation of decision-making, which leads to the user's responsibility. However, in cases where the interaction between generative AI and humans is so dense that individual contributions cannot be identified,[61] responsibility would be determined according to the principles of network liability. Finally, the socialization of risks via collective funds should be considered only for technologies that operate below the societal level in an interconnected digital sphere without direct interaction with humans.

61 Mark Coeckelbergh and David J Gunkel, 'ChatGPT: Deconstructing the Debate and Moving it Forward' (2024) *AI & Society* 2221, 2225; Joerge Luis Morton Gutiérrez, 'On Actor-Network Theory and Algorithms: ChatGPT and the New Power Relationships in the Age of AI' (2024) *AI and Ethics* 1071, , 1077 ff.