Enterprise Liability for Human-AI-Decisions:

A Multidisciplinary Approach for Identifying Principals' Duties of Care

Ann-Kristin Mayrhofer*

Contents

* Ann-Kristin Mayrhofer is Akademische Rätin a. Z. (Research Fellow) at the Chair of Civil Law, Civil Procedure, European Private Law and Procedure of Prof. Dr. Beate Gsell at Ludwig-Maximilians-Universität München.

University of Vienna Law Review, Vol. 9 No 3 (2025), pp. 35-63, https://doi.org/10.25365/vlr-2025-9-3-35.

I. Introduction

Artificial intelligence (AI)[1] systems are already outperforming humans in a significant number of tasks, such as diagnosing diseases, classifying objects in images or predicting loan default risks,[2] and it is likely that this number will keep rising.[3] AI systems typically excel over humans when the task involves processing vast amounts of data and making general assumptions.[4] At the same time, AI systems are still limited in ways humans are not. They usually have difficulties when the task requires them to deviate from a general pattern and take into account the particularities of an individual case.[5] The following example, concerning driving, illustrates the respective strengths and weaknesses of AI systems and humans: An AI driver confronted with

[1] In this paper, the – controversial – term of "Artificial Intelligence" is understood in a broad sense: It is meant to include "simple" algorithms which are not based on machine learning (ML) techniques; cf. the definition of "AI system" in Art. 3(1) of the Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (AI Act): "a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments".

[2] Cf. Yotam Liel and Lior Zalmanson, 'Turning Off Your Better Judgment – Conformity to Algorithmic Recommendations' (Working paper) (2022), available at https://www.researchgate.net/publication/366412145_Turning_Off_Your_Better_Judgment_-Conformity_to_Algorithmic_Recommendations (last accessed 25 April 2025), pp. 5-6 (with further references; an abbreviated version of this working paper was published in Academy of Management Proceedings 2023(1); hereafter only the extended working paper will be referenced).

[3] For an overview of use cases, cf. Deutsches Institut für Normung and Deutsche Kommission Elektrotechnik, Elektronik, Deutsche Normungsroadmap Künstliche Intelligenz, 2nd edn. (2022), available at https://www.din.de/de/forschung-und-innovation/themen/kuenstliche-intelligenz/fahrplan-festlegen (last accessed 25 April 2025); Katja Grace et al., 'Viewpoint: When Will AI Exceed Human Performance? Evidence from AI Experts' (2018) 62 Journal of Artificial Intelligence Research 729-54 conducted a survey among ML researchers in 2016 and found that, on average, they believed there was a 50 % chance AI would outperform humans in all tasks in 45 years' time.

[4] Cf. Kathleen L. Mosier and Linda J. Skitka, 'Human Decision Makers and Automated Decision Aids: Made for Each Other?', in Raja Parasuraman and Mustapha Mouloua (eds.), Automation and human performance: Theory and application (Mahwah: Lawrence Erlbaum, 1996) 201–20, pp. 201 and 209.

[5] Cf. Kevin Bauer et al., 'Die KI braucht bei der Bankberatung immer noch menschliche Hilfe', Börsen-Zeitung, available at https://www.boersen-zeitung.de/kapitalmarktforschung/in-der-bankberatung-braucht-die-ki-menschliche-hilfe-90edbb42-86a4-11ed-a311-f90ecc32c8e4 (last accessed 25 April 2025); for a philosophical view on AI-decision-making, cf. Andreas Kaminski, 'Gründe geben. Maschinelles Lernen als Problem der Moralfähigkeit von Entscheidungen', in Klaus Wiegerling et al. (eds.), Datafizierung und Big Data (Wiesbaden: Springer VS, 2020) 151-74.

University of Vienna Law Review, Vol. 9 No 3 (2025), pp. 35-63 https://doi.org/10.25365/vlr-2025-9-3-35.

the usual road signs may cause less accidents than a human driver. However, if a new sign appears, e.g., in a new country, which does not match the system's knowledge, the AI driver could cause an accident a human driver may have avoided by stopping and asking a pedestrian about the sign's meaning.[6] Furthermore, small changes to traffic signs, that are invisible to the human eye and do not change the signs' meaning, could fool AI systems and lead to accidents humans would not have caused.[7] Generally, there are some damage risks that are better avoided by AI systems and others that only humans could help to prevent.[8] This suggests that "human–ML [machine learning] augmentation, where humans and technology work together to perform organisational tasks jointly" may be "the most promising path".[9] In fact, it is expected that enterprises will increasingly integrate Human-AI-Decisions into their organisations.[10] In principle, a Human-AI-Decision could replace any decision previously made by either a human or an AI system. It could, for example, concern medical diagnoses,[11] credit scoring[12] or the safety of products or services offered by the enterprise[13]. Combining humans and AI systems could enable the utilisation of both human and AI potential and therefore increase efficiency and safety. However, while many damage risks may be reduced by this combination, the Human-AI

[6] Cf. the similar example by Erik J. Larson, The myth of artificial intelligence, p. 124.

[7] Cf. Kevin Eykholt et al., 'Robust Physical-World Attacks on Deep Learning Visual Classification', in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Salt Lake City: IEEE, 2018) 1625-34.

[8] Cf. Philipp Hacker, 'Verhaltens- und Wissenszurechnung beim Einsatz von Künstlicher Intelligenz' (2018) 9(3) RW 243-88, p. 263 who distinguishes between three types of errors: errors which are only committed by AI systems, errors which are committed both by (reasonable) humans and AI systems and errors which are only committed by humans.

[9] Cf. Mike H. M. Teodorescu et al., 'Failures of Fairness in Automation Require a Deeper Understanding of Human-ML Augmentation' (Minneapolis: University of Minnesota 2021) 45(3) MIS Quarterly 1483-500, p. 1484 on "achieving fairness".

[10] Cf. Liel and Zalmanson, 'Turning Off Your Better Judgment', p. 2 (with further references).

[11] Cf. Ekaterina Jussupow et al., 'Augmenting Medical Diagnosis Decisions? An Investigation into Physicians' Decision-Making Process with Artificial Intelligence' (2021) 32(3) Information System Research 713-35.

[12] Cf. Rita Gsenger and Toma Strle, 'Trust, Automation Bias and Aversion: Algorithmic Decision-Making in the Context of Credit Scoring' (2021) 19(4) Interdisciplinary Description of Complex Systems 542-60.

[13] For a concept of an autonomous warehouse, see e.g., Ahmet Börütecene and Jonas Löwgren, 'Designing Human-Automation Collaboration for Predictive Maintenance', in *Companion Publication of the 2020 ACM Designing Interactive Systems Conference* (New York: Association for Computing Machinery, 2020) 251-6.

cooperation may also give rise to new damage risks.[14] These risks trigger the question of liability. The most obvious issue seems to be the liability of the human who directly cooperates with the AI system. However, this paper addresses the liability of the *enterprise* that makes use of the Human-AI cooperation. More precisely, it is the liability of the *principal*, the corporation or the single entrepreneur that holds the enterprise, which is at stake.[15] Compared to the human agent's liability, enterprise liability may provide considerable advantages for victims: Usually, it is easier to identify the enterprise the Human-AI-Decision is integrated in than the individual human cooperating with the AI system. Furthermore, the principal is usually in a better financial situation.[16]

This paper focuses on the enterprise's *non-contractual* and *fault-based* liability.[17] Its aim is to use findings from other disciplines to identify some of the duties of care that principals must comply with. The paper will illustrate the need for a multidisciplinary approach in the context of technology particularly regarding Human-AI-Decisions[18] whose facets "are as complex as the environments in which they function".[19] Given the author's background, the analysis is based on a *German perspective*. However, it sems that the general ideas can also be applied to other legal systems. The paper first lays some principles of fault-based liability law and, at the same time, sets out the methodological framework of the applied multidisciplinary approach (II.). It will be shown that the notions of "possibility" and "reasonableness", commonly used to define duties of care, can serve as a methodological gateway to such approach.

---

[14] Cf. the examples at Mosier and Skitka, 'Human Decision Makers and Automated Decision Aids', pp. 205-6.

[15] Non-human principals, namely corporations, generally act through human representatives (see, for example, § 31 of the German Civil Code (BGB)).

[16] Cf. Helmut Koziol, 'Concluding Remarks', in Helmut Koziol (ed.), *Comparative Stimulations for Developing Tort Law* (Vienna: Sramek, 2015) 182-95, paras 5-27.

[17] For a broader analysis of liability for autonomous systems (humans, animals, and AI systems), cf. Ann-Kristin Mayrhofer, *Außervertragliche Haftung für fremde Autonomie* (Tübingen: Mohr Siebeck, 2023).

[18] Cf. Anna Beckers and Gunther Teubner, *Three Liability Regimes for Artificial Intelligence* (Oxford: Hart, 2022), p. 16. One may argue whether the following approach is "multidisciplinary" or already "interdisciplinary" (or "transdisciplinary"). The borders are fluid, and the terminology does not influence the following considerations; for definitions, see Benard C.K. Choi and Anita W.P. Pak, 'Multidisciplinarity, interdisciplinarity, and transdisciplinarity in health research, services, education and policy: 1. Definitions, objectives, and evidence of effectiveness' (2006) 29(6) Clinical and Investigative Medicine 351-64; Jacqueline Fawcett, 'Thoughts About Multidisciplinary, Interdisciplinary, and Transdisciplinary Research' (2013) 26(4) Nursing Science Quarterly 376–9; Eric Hilgendorf, 'Bedingungen gelingender Interdisziplinarität' (2010) 65(19) JZ 913-22, pp. 914-5.

[19] Mosier and Skitka, 'Human Decision Makers and Automated Decision Aids', p. 202.

University of Vienna Law Review, Vol. 9 No 3 (2025), pp. 35-63 https://doi.org/10.25365/vlr-2025-9-3-35.

Second, the paper examines – as its main part – measures which have been identified by non-legal experts as "possible" mitigators of risks associated with Human-AI-Decisions (III.). Third, the paper explores – more briefly – how economic considerations may help to limit principals' duties to the "reasonable" measures (IV.).[20] Finally, it looks at the European Commission's Proposal for an AI Liability Directive and its significance for Human-AI-Decisions and the principals' corresponding duties of care (A.). The AI Liability Directive was proposed in 2022 and abandonded in 2025.

II. Fault-Based Liability and Principals' Duties of Care: The Notions of "Possibility" and "Reasonableness" as Methodological Gateway to Multidisciplinary Approaches

Most legal systems recognise various forms of non-contractual liability.[21] This paper focuses on *fault-based* liability and more specifically on negligence liability. Fault-based liability "has been recognised for centuries as an element of liability"[22] and, unlike e.g., strict liability,[23] applies in all kinds of situations. Comparative studies have shown that the term of "fault" could refer to a variety of different concepts.[24] Yet, there seems to be a certain agreement that negligence liability presupposes at least the violation of a *duty of care*.[25] Therefore, this paper seeks to identify some of the principals' duties of care in relation to *Human-AI-Decisions* integrated in their

[20] For an economic analysis of AI liability, see Gerhard Wagner, 'Roboter als Haftungssubjekte? Konturen eines Haftungsrechts für autonome Systeme', in Florian Faust and Hans-Bernd Schäfer (eds.), *Zivilrechtliche und rechtsökonomische Probleme des Internet und der künstlichen Intelligenz* (Tübingen: Mohr Siebeck, 2019) 1-39.

[21] For an overview of tort law in different jurisdictions, cf. Helmut Koziol (ed.), *Basic Questions of Tort Law from a Germanic Perspective* (Vienna: Sramek, 2012); Helmut Koziol (ed.), *Basic Questions of Tort Law from a Comparative Perspective* (Vienna: Sramek, 2015) and Helmut Koziol (ed.), *Comparative Stimulations for Developing Tort Law* (Vienna: Sramek, 2015).

[22] Helmut Koziol, 'Comparative Conclusions' in Helmut Koziol (ed.), *Basic Questions of Tort Law from a Comparative Perspective* (Vienna: Sramek, 2015) 685-838, para 8/218.

[23] With strict liability "European legal systems show much more diversity than in other areas of tort law", ibid., para 8/35.

[24] Cf. ibid., paras 8/219-224; Jan De Bruyne et al., 'The European Commission's approach to extra-contractual liability and AI – An evaluation of the AI liability directive and the revised product liability directive' (2023) 51 Computer Law & Security Review 105894, pp. 8-9.

[25] Cf. Koziol, 'Comparative Conclusions', in Koziol (ed.), *Basic Questions of Tort Law from a Comparative Perspective*, paras 8/224 and 8/226. However, there is no complete consensus on this concept, cf. Gert Brüggemeier, *Haftungsrecht: Struktur, Prinzipien, Schutzbereich* (Berlin et al.: Springer, 2006), pp. 52-5 (with further references); De Bruyne et al., 'The European Commission's approach to extra-contractual liability and AI', p. 9 (with further references).

organisations. The practical significance of principals' duties of care for the enterprise's liability varies between jurisdictions:[26] Under German law, for example, principals are only liable for damages caused by their – human or artificial – assistants if the principals acted faultily *themselves*, e.g., by violating their supervision duties.[27] Other jurisdictions, for example the French system, contain stricter rules.[28] In these jurisdictions, the question of whether a duty of care has been violated is less pressing. Liability for Human-AI-Decisions under such specific and, to a significant extent, system dependent rules will not be discussed in this paper either.[29]

Principals who integrate humans into their organisations must carefully select, instruct, and supervise *each person*.[30] In principle, the same applies to technical devices, including AI systems.[31] However, duties of care are not limited to the direct relationship between principals and – human or artificial – assistants. Rather, principals also need to coordinate the cooperation *between* the assistants.[32] Regarding German law, the German Federal Court of Justice (Bundesgerichtshof, BGH) holds that a person does not need to avoid *every* risk but must (only) do what is "possible" and "reasonable".[33] As the idea that "no one is obliged beyond what they are able to do" (*ultra posse nemo obligatur*) and the standard of the "reasonable person" are

[26] For a comparative overview cf. Koziol, 'Comparative Conclusions', in Koziol (ed.), *Basic Questions of Tort Law from a Comparative Perspective*, paras 8/250-5 and the contributions in Helmut Koziol et al., 'Liability for Agents and Agents' Liability', in Helmut Koziol (ed.), *Comparative Stimulations for Developing Tort Law* (Vienna: Sramek, 2015) 182-95, pp. 147-95.

[27] Cf. §§ 831, 823 BGB. However, § 831 BGB contains a presumption of the violation of a duty of care.

[28] Cf. Art. 1242 of the French Civil Code.

[29] Therefore, the question of whether specific principal-agent-liability applies to AI systems will also remain open. At least, in more restrictive systems, such as the German system, an analogy would not lead to principals' strict or vicarious liability, for the discussion, cf. Hacker, 'Verhaltens- und Wissenszurechnung', pp. 265-7; Herbert Zech, 'Entscheidungen digitaler autonomer Systeme: Empfehlen sich Regelungen zu Verantwortung und Haftung?', in Ständige Deputation des Deutschen Juristentages (ed.), *Verhandlungen des 73. Deutschen Juristentages* (Munich: C.H. Beck, 2020) A 1-112, pp. A 76-81 and A 95-96; Beckers and Teubner, *Three Liability Regimes for Artificial Intelligence*, pp. 46-87 and passim.

[30] Cf. § 831(1) sentence 2 BGB.

[31] Cf. Susanne Horner and Markus Kaulartz, 'Haftung 4.0' (2016) 32(1) CR 7-14, p. 8.

[32] Mayrhofer, *Außervertragliche Haftung für fremde Autonomie*, p. 111 and 343 with further references regarding so-called "organisational duties" ("Organisationspflichten").

[33] BGH, 6 February 2007, VI ZR 274/05 (2007) NJW 1683-5, para 14; cf., also, Gerhard Wagner, '§ 823 BGB', in Jürgen Säcker et al. (eds.), *Münchener Kommentar zum Bürgerlichen Gesetzbuch*, 9th edn., 13 vols. (Munich: C.H. Beck, 2024), vol. VII, para. 528 (with further references).

largely recognised in most national laws,[34] it seems that the duties of care are defined similarly in other jurisdictions.

In simple cases the required safety standard may be obvious. For instance, it is certainly "possible" and "reasonable" to slow down before a sharp turn to avoid accidents. Difficulties arise in more complex cases where assessing the "possibility" and "reasonableness" of a measure requires a lot of non-legal knowledge. This is where multidisciplinary approaches can be useful:[35] Such approaches can help to concretise abstract concepts,[36] such as "possibility" and "reasonableness"[37]. These open-formulated requirements can serve as a methodological "gateway" to multidisciplinary approaches.[38] Cases of damages that involve Human-AI-Decisions tend to be complex: Humans and AI systems are complex entities, and the complexity increases when they work together.[39] Determining whether a particular measure provides a safety benefit and whether that benefit outweighs the effort that a principal must take to adopt the measure, is very difficult.[40]

Written standards, such as product safety legislation and technical standards, can provide some relief. However, they usually do not cover every situation exhaustively,

[34] Both the idea that "no one is obliged beyond what they are able to do" (*ultra posse nemo obligatur*) and the standard of the "reasonable person" seem to be largely recognised in most national laws, cf. on the latter Art. 4:102 (1) Principles of European Tort Law (PETL), available at http://www.egtl.org/PETLEnglish.html (last accessed 25 April 2025); Koziol, 'Comparative Conclusions', in Koziol (ed.), *Basic Questions of Tort Law from a Comparative Perspective*, paras 8/229-36.

[35] Cf. Franz Hofmann, 'Disziplinarität, Interdisziplinarität und Interdisziplinarität am Beispiel der Grundsätze „mittelbarer Verantwortlichkeit"' (2018) 73(15-16) JZ 746-54, pp. 749-50 and 753-4.

[36] Cf. Hilgendorf, 'Bedingungen gelingender Interdisziplinarität', p. 920.

[37] Cf. Hofmann, 'Disziplinarität, Intradisziplinarität und Interdisziplinarität', pp. 750 and 753-754 who uses interdisciplinary approaches to identify duties of care in the form of so called "Verkehrspflichten".

[38] Cf. ibid., p. 750 ("Einfallstor"); Rolf Stürner, 'Die Zivilrechtswissenschaft und ihre Methodik – zu rechtsanwendungsbezogen und zu wenig grundlagenorientiert?' (2014) 214(1-2) AcP 7-54, pp. 31-32 ("Einbruchstellen").

[39] Cf. Expert Group on Liability and New Technologies, 'Liability for Artificial Intelligence and Other Emerging Digital Technologies' (2019), available at https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/JURI/DV/2020/01-09/AI-report_EN.pdf (last accessed 25 April 2025), pp. 32-3.

[40] Cf. in the context of product liability, Mayrhofer, *Außervertragliche Haftung für fremde Autonomie*, p. 269.

especially when it comes to rapidly evolving technologies such as AI.[41] The AI Act,[42] for example, will certainly play an important role in the assessment of fault-based liability. It focuses mainly on "high-risk" AI systems and "general-purpose AI models".[43] Yet, many of its requirements leave room for interpretation,[44] such as the need for high-risk AI systems to be "effectively overseen by natural persons".[45] Thus, the written standards will most likely need to be supplemented by courts which will need to equip themselves with the non-legal knowledge necessary to adequately assess the specific risks of the Human-AI-Decision. In theory, the need for multidisciplinary approaches seems to be recognised by the courts. The BGH, for example, refers to the "state of scientific and technical knowledge" when determining whether a safety measure was "possible".[46] According to the BGH, enterprises are able to avoid a risk "if, according to the assured expert knowledge of the relevant specialist groups, solutions are available that can be used in practice".[47] Furthermore, when determining, what is "reasonable", the BGH considers "all the circumstances of the individual case", in particular "the magnitude of the risk", but also "the economic effects of the safeguarding measure".[48] This wording suggests that the safety standards are indeed set using knowledge from other disciplines, including technical, sociological and economic aspects. In practice, however, it can be difficult for judges to make use of non-legal knowledge.[49] Therefore, the paper will now show how studies on Human-AI-Decisions can effectively be made fruitful for the legal analysis of the principals' duties of care. It will present examples of non-legal studies and draw some conclusions on the legal assessment of "possibility" and "reasonableness". While the first step – presentation of the non-legal knowledge – is rather descriptive,

[41] Cf. in the context of product liability, ibid., pp. 275-85.

[42] Cf. note 1 above.

[43] Principals that integrate Human-AI-Decisions into their organisations will frequently qualify as "deployer" under the AI Act (Art. 3(4)). However, if they develop the AI system or general-purpose AI model themselves or have the AI system developped, they can also be classified as "providers" (Art. 3(3)).

[44] Cf. David Bomhard and Marieke Merkle, 'Europäische KI-Verordnung – Der aktuelle Kommissionsentwurf und praktische Auswirkungen' (2021) 1(6) RDi 276-83, p. 283.

[45] Cf. Art. 14 AI Act.

[46] BGH, 16 September 2009, VI ZR 107/08, BGHZ 181, 253-68, para 16 (juris). The decisions mainly concern manufacturers' liability. However, there is no reason for limiting this approach to such enterprises so that it may also apply to e.g., service providers.

[47] Ibid.

[48] Ibid., para 18.

[49] Cf. Hilgendorf, 'Bedingungen gelingender Interdisziplinarität', p. 917.

the second step – legal conclusions – aims to provide some new guidelines for the assessment of duties of care.

III. "Possible" Measures: Specific Risks of Human-AI-Decisions and Risk Mitigation

Human-AI cooperation can take various forms. Three categories are frequently employed: First, AI systems can *provide information* to the human. Second, they can *make recommendations*. Third, AI systems can *take the decision* on behalf of the human.[50] This paper focuses on AI systems that generate an output that the human can *accept* or *reject* before it causes damage. The human is not completely replaced by the AI system but stays "in the loop".[51] Mostly, these systems give recommendations (second category). However, the output could also consist of information the human can disregard (first category) or a decision the human can override (third category).[52]

The specific risks of Human-AI-Decisions explored here are risks associated with precisely this option to *accept* or *reject* an AI output. The aim of Human-AI cooperation – utilising both AI and human strengths – is only achieved if humans correctly "exercise their own judgment [...] to minimise risks of bad or biased decisions".[53] Otherwise, correct AI output may be rejected or wrong AI output may be accepted, and damages may occur, also to third parties. Particularly, just as when working with other humans, humans need to have the right level of trust and mistrust

[50] Cf. the definition of an "AI system" in Art. 3(1) AI Act: "generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments"; Datenethikkommission der Bundesregierung, *Gutachten* (2019), available at https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.html (last accessed 25 April 2025), pp. 161-2: "algorithm-based", "algorithm-driven" and "algorithm-determined and therefore completely automated" decisions.

[51] Cf. High-Level Expert Group on Artificial Intelligence, 'Ethics Guidelines for Trustworthy AI' (2019), available at https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (last accessed 25 April 2025), p. 16, where a distinction is made between three approaches of human oversight: "human-in-the-loop (HITL), human-on-the-loop (HOTL), or human-in-command (HIC)" and where HITL is defined as "the capability for human intervention in every decision cycle of the system".

[52] Cf. the "Levels of Automation of Decision and Action Selection" in Matthew Ball and Vic Callaghan, 'Explorations of Autonomy', in *2012 Proceedings of the 8th International Conference on Intelligent Environments* (Guantajo: IEE Xplore, 2012) 114-21.

[53] Cf. Liel and Zalmanson, 'Turning Off Your Better Judgment', p. 2 (with further references).

towards their artificial "colleague".[54] However, three major phenomena appear to hinder humans from adopting the appropriate attitude towards AI systems, which can be categorized into two groups:[55] (A.) Automation Complacency and Automation Bias on the one hand and (B.) Algorithmic Aversion on the other. Non-legal experts have conducted research on these problems and have suggested mitigating factors that can help to determine the "state of scientific and technical knowledge" regarding Human-AI-Decisions and to identify "possible" risk prevention measures.

A. Automation Complacency and Automation Bias: Over-Reliance on AI Systems

In the context of this paper, Automation Complacency and Automation Bias can be addressed together. They "represent different manifestations of overlapping automation-induced phenomenon"[56] as they both lead to "inappropriate overreliance on automation".[57] While Automation Complacency is generally linked to insufficient monitoring of a system's output,[58] Automation Bias describes humans' tendency to accept the output "as a heuristic replacement for vigilant information seeking and processing".[59] Research on these phenomena has started in the aviation sector: Human pilots frequently did not monitor the autopilots properly or accepted sub-optimal flight plans.[60] Subsequently, many studies have been conducted in different contexts to find out more about causes and possible mitigators of human over-

[54] Cf. Gsenger and Strle, 'Trust, Automation Bias and Aversion', pp. 545-6.

[55] There seem to be other phenomena, e.g., "selective adherence", cf. Saar Alon-Barkat and Madalina Busuioc, 'Human-AI Interactions in Public Sector Decision-Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice' (2023) 33(1) Journal of Public Administration Research and Theory 153–69, p. 154.

[56] Raja Parasuraman and Dietrich H. Manzey, 'Complacency and Bias in Human Use of Automation: An Attentional Integration' (2010) 53(3) Human Factors 381-410, p. 405.

[57] Ibid., p. 398.

[58] Ibid., p. 382; the authors point out that "there is no consensus on the definition of complacency". Also, cf. Kate Goddard et al., 'Automation bias: a systematic review of frequency, effect mediators, and mitigators' (2012) 19(1) Journal of the American Medical Informatics Association 121-7, pp. 121-2.

[59] Mosier and Skitka, 'Human Decision Makers and Automated Decision Aids', p. 205; cf. Goddard et al., 'Automation bias', p. 121.

[60] Cf. Gsenger and Strle, 'Trust, Automation Bias and Aversion', p. 546; Parasuraman and Manzey, 'Complacency and Bias in Human Use of Automation', pp. 381-2; Mary L. Cummings, 'Automation Bias in Intelligent Time Critical Decision Support Systems' (2004) AIAA 1st Intelligent Systems Technical Conference, 20-22 September 2004, Chicago, IL, American Institute of Aeronautics and Astronautics 1-6 (with reference to further studies).

University of Vienna Law Review, Vol. 9 No 3 (2025), pp. 35-63 https://doi.org/10.25365/vlr-2025-9-3-35.

reliance on algorithms.[61] Not all of them can be examined here. Rather, this paper focuses on two examples dealing with two distinct kinds of decisions.

1. "Difficult" Decisions: Medical Diagnostic

Jussupow et al. conducted an experiment with physicians who had to make medical diagnoses based on radiological data and advice provided by an AI system.[62] In the control group – physicians without AI advice – the accuracy rate, manifesting the decision-making *results*, was about 77 %. The authors found that the accuracy rate was significantly lower among the physicians who received incorrect AI advice (about 55 % accuracy rate).[63] To investigate the decision-making *process*, they also analysed think-aloud protocols, questionnaires and interviews.[64] According to *Jussupow* et al., three "decision pathways" could lead to erroneous conformity to AI advice: First, physicians tend to rely on AI advice for their own *initial* assessment. Second, when AI advice *disconfirms* their initial position, novice physicians are likely to doubt their own capabilities. They lean heavily on their *belief* in the *system* without further *validation* of the specific *assessments* ("belief conflict" vs. "validation conflict"). Experienced physicians more often ignore the AI advice or engage in such an evaluation of the advice. However, in case of such validation, physicians frequently do not reconsider both their own *and* the AI assessment but only collect data to reject or confirm *one* of the positions. This leads to the third pathway: If physicians opt for only checking the AI assessment, they often fail to find or even search for data that contradicts the AI advice.[65]

Regarding mitigators against erroneous conformity to AI advice, the authors suggest that the system *design* could have an influence. Physicians seem to ignore wrong advice more often when it is presented *after* they have already made an initial assessment.[66] This idea is in line with findings of behavioural economists who suggest improving the decisions of (human) groups by having participants form their opinions

[61] Cf. III.A.1. and III.A.2. below, and the examples provided by Liel and Zalmanson, 'Turning Off Your Better Judgment', pp. 3-8.

[62] Jussupow et al., 'Augmenting Medical Diagnosis Decisions?'.

[63] Ibid., pp. 721-2. In contrast, the accuracy rate was only marginally higher than the accuracy rate of the control group among those participants who received correct advice (about 90 % accuracy rate).

[64] Ibid., p. 718.

[65] Ibid., pp. 729-30 and passim; cf. already Mosier and Skitka, 'Human Decision Makers and Automated Decision Aids', p. 202-205 where humans' tendency to interpret other information as being consistent with the automated decision aid is highlighted.

[66] Jussupow et al., 'Augmenting Medical Diagnosis Decisions?', p. 730.

before the discussion.[67] *Jussupow* et al. also highlight the importance of training: Physicians need to learn how to validate the AI system's *and* their own position, so that first, they do not get stuck in the "belief conflict" and second, they are able to solve the "validation conflict" by reconsidering both their own[68] *and* the AI's assessment ("system monitoring" and "self-monitoring").[69]

2. "Easy" Decisions: Simple Image Classification

Liel and *Zalmanson* cite *Jussupow* et al. as an example of a context "in which the purpose of the algorithm was to assist in complex calculations or judgment under uncertainty".[70] According to *Liel* and *Zalmanson*, in such situations, erroneous conformity to AI advice could be "justifiable".[71] Their own experiment, by contrast, dealt with "objective and straightforward tasks, with very little uncertainty".[72] Participants were shown a set of three items alongside a fourth one and were asked to identify the item in the set which was equal in length to the fourth item. The participants were split into three groups: A control group which received no advice, a second group which received wrong advice framed as being generated by an AI, and a third group which received wrong advice framed as being generated by humans.[73] Members of the control group were accurate in around 98 % of the tasks which demonstrates the tasks' simplicity. Members of the second group followed the erroneous AI advice in around 27 % of the tasks, whereas members of the third group followed the erroneous human advice in about 19 % of the tasks.[74]

According to *Liel* and *Zalmanson*, these results show that humans do not only over-rely on AI advice in cases of uncertainty. Rather, they also exhibit an "'irrational'

[67] Cf. Daniel Kahneman et al., 'NOISE' (2016) 94(10) Harvard Business Review 38-46, p. 46.

[68] Cf. also Andreas Fügener et al., 'Cognitive Challenges in Human–Artificial Intelligence Collaboration: Investigating the Path Toward Productive Delegation' (2022) 33(2) Information Systems Research 678–96: The authors conducted experiments where humans did not perform well in delegating tasks to AI systems. They explain this result mainly by humans' lack of knowledge about their own capacities ("metaknowledge").

[69] Jussupow et al., 'Augmenting Medical Diagnosis Decisions?', pp. 731-2.

[70] Liel and Zalmanson, 'Turning Off Your Better Judgment', p. 7.

[71] Ibid., p. 8.

[72] Ibid., p. 8.

[73] Ibid., pp. 15-8.

[74] Ibid., pp. 18-20.

tendency to adopt advice that contradicts their better judgment".[75] The comparison with human advice indicates that at least in certain situations "algorithmic conformity" can even exceed "social conformity".[76] One could argue that, in real life, such simple tasks are unlikely to be carried out by Human-AI cooperation.[77] However, it seems that the findings could be transferred to AI advice concerning more difficult tasks, e.g., driving. AI advice could also be obviously wrong, e.g., if the AI driver clearly misinterprets a new sign. Adherence to such AI advice can very well have serious real-life impact. This hypothesis is supported by anecdotes about accidents involving clearly erroneous navigation system.[78]

Liel and *Zalmanson* also investigated factors that could mitigate conformity. They conducted more experiments and found that the tendency to conform to incorrect AI advice was significantly lower when a second AI system provided *correct* advice.[79] In addition, participants seemed to perform better when they were told that their task was of "high significance", e.g., that it was to improve the safety of autonomous vehicles.[80] Consequently, they suggest including two or more AI systems in a Human-AI-Decision and encouraging users to perceive their tasks as important.[81]

B. Algorithm Aversion: Under-Reliance on AI Systems

The aforementioned studies highlight that AI systems may give wrong advice in some cases. However, as AI systems frequently outperform humans, in other cases, humans should better trust the output of AI systems. In these situations, damages occur if humans do *not* use AI advice but rely on their own initial judgment or the

[75] Ibid., p. 33.

[76] The authors explicitly refer to experiments on social conformity. Their experimental design is based on studies conducted by *Solomon E. Asch* in the 1950s, cf. ibid., pp. 16-8.

[77] Ibid., p. 8.

[78] Cf. e.g., Greg Milner, 'Death by GPS: Are Satnavs changing our brains?', *The Guardian*, available at https://www.theguardian.com/technology/2016/jun/25/gps-horror-stories-driving-satnav-greg-milner (last accessed 25 April 2025). Reference to such "anecdotal evidence" of automation bias is also made by Alon-Barkat and Busuioc, 'Human-AI Interactions in Public Sector Decision-Making', p. 155.

[79] Liel and Zalmanson, 'Turning Off Your Better Judgment', pp. 20-4.

[80] One group was told the work served to improve the safety in autonomous vehicles, one group was told that it served to improve performance in smart warehouses, and one group did not receive any context, cf. ibid., pp. 24-7.

[81] Ibid., pp. 34-5.

advice of other humans. This phenomenon is called Algorithm Aversion.[82] To explore the risks associated with this third problem and possible mitigators, again, two studies will be presented:

The first study was conducted by *Dietvorst* et al. Participants were provided with data about MBA students or U.S. states. They were then asked to predict the students' performance in the MBA program or the rank of U.S. states in terms of the number of airline passengers departing from that state.[83] Participants had to choose between two advisors: a human or a statistical model. Prior to their predictions, participants either saw the AI system's performance on this task, the human's performance, both or neither.[84] The results suggest that Algorithm Aversion is particularly prevalent when humans have seen AI systems performing and making mistakes: Even after seeing the AI system outperforming the human, participants preferred human advisors.[85] Therefore, it seems that "resistance at least partially arises from greater intolerance for error from algorithms than from humans".[86] At the same time, this finding suggests that Algorithm Aversion may be prevented by hiding the AI system's past errors from the human or by hiding the AI system itself.[87]

The second study, conducted by *Yeomans* et al., dealt with selecting jokes.[88] The authors found that AI systems generally outperformed humans in this task. They were usually able to select jokes which participants found funnier.[89] Nonetheless, participants seemed to prefer receiving recommendations from humans: They rated the advisor better when they believed it was human.[90] *Yeomans* et al. also found that advice that was framed as coming from a machine was perceived as less scrutable than advice framed as coming from a human.[91] To the authors, this indicates the

[82] Cf. Berkeley J. Dietvorst et al., 'Algorithm aversion: People erroneously avoid algorithms after seeing them err' (2015) 144(1) Journal of Experimental Psychology: General 114-26, p. 114.

[83] Ibid., pp. 115 and 118.

[84] Ibid., p. 115.

[85] Ibid., p. 119.

[86] Ibid., p. 124.

[87] Cf. Ibid., p. 124.

[88] Michael Yeomans et al., 'Making sense of recommendations' (2019) 32(4) Journal of Behavioral Decision Making 403-14.

[89] Ibid., pp. 404-8.

[90] Ibid., pp. 409-10.

[91] Ibid., pp. 410-1.

"subjective understanding" of the recommenders influence the assessment.[92] This idea was supported in another experiment, where one group received only a sparse explanation of the system's decision-making process whereas the other one got details about the functioning.[93] It suggests that "algorithmic sensemaking" could mitigate Algorithm Aversion.[94] Such sensemaking could, for example, be achieved by improving the AI system's explainability.[95] *Yeomans* et al. themselves propose that more experience with the AI system could increase humans' subjective understanding.[96] However, more experience with the system may be accompanied by seeing the system making more mistakes which, according to the aforementioned studies of *Dietvorst* et al., could foster Algorithm Aversion. According to *Yeomans* et al, another way of increasing subjective understanding could be creating AI systems with a more human-like design.[97] Yet, it seems that strong human resemblance could equally have negative effects on humans' attitude toward the AI system.[98] Besides, such a design could make humans belief that the AI system takes its decisions the same way a human does. This is usually wrong and could equally lead to a false re-evaluation of the advice.[99]

C. Conclusions: Which Measures are "Possible"?

The research just presented might assist in identifying "possible" measures to avoid the risks of Human-AI-Decisions. Naturally, the extent to which Automation Complacency, Automation Bias and Algorithm Aversion influence Human-AI-Decision-making depends on the specific task.[100] For example, a study by *Lee* which dealt with managerial decisions suggests that when a task requires mostly

[92] Ibid., pp. 410-1.

[93] Ibid., p. 411.

[94] Ibid., p. 412.

[95] Cf. on transparency measures, Fraunhofer IAIS, *Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz* (2021), available at https://www.iais.fraunhofer.de/de/publikationen/studien/2021/ki-pruefkatalog.html (last accessed 25 April 2025), pp. 63-85.

[96] Michael Yeomans et al., 'Making sense of recommendations', p. 412.

[97] Ibid., p. 412.

[98] Cf. the idea of an "uncanny valley" developed by the Japanese professor *Mashiro Mori* in the 1970s, Mashiro Mori, 'The Uncanny Valley' (2012) 19(2) IEEE Robotics & Automation Magazine 98–100 (translated by Karl F. MacDorman and Norri Kageki).

[99] Cf. already Mosier and Skitka, 'Human Decision Makers and Automated Decision Aids', pp. 210.

[100] Cf. Gsenger and Strle, 'Trust, Automation Bias and Aversion', p. 548.

"mechanical" skills, e.g., work scheduling, AI and human advice appear equally trustworthy, whereas in more "human" tasks, e.g., work evaluation, AI advice is perceived less trustworthy.[101] *Alon-Barkat* and *Busuioc* conducted a study on school board decisions concerning the employment of teachers and found no evidence for Automation Bias.[102] According to the authors one reason might be a "relative skepticism about the performative capacity of AI algorithms" which distinguished this task from "areas well-accustomed to such devices (aviation, healthcare), characterised by routine use of reliable automation, resulting in high levels of trust in their performance".[103] *Liel* and *Zalmanson* point out that the situation they experimented with – image classification – "may be characterised by particularly high levels of trust in technology and algorithms",[104] which could certainly encourage conformity. However, despite these differences and difficulties, the studies and their interpretation by their authors allow drawing some general conclusions regarding "possible" risk prevention measures.

Some of these measures concern the *design* of the Human-AI cooperation:[105] Principals can ensure humans make an initial assessment *before* receiving the AI advice, e.g., by making them write down their own opinion first (cf. III.A.1.). Besides, they can provide human agents with advice from *multiple* AI systems (cf. III.A.2.). Additionally, *training* could serve as a mitigator: Humans can be taught how *both* AI systems *and* humans make their decisions so that in case of a disagreement they are able to re-evaluate *both* the AI's advice *and* their own initial assessment (cf. III.A.1.). Furthermore, regarding specific task, *information* plays a significant role: Principals can inform the human agents about the *general performance* of specific AI systems and, if a comparison is possible, about the performance of competing humans. They can clarify to the human the *importance* of the task (cf. III.A.2.) and the *organisation's expectations*.[106] Besides, principals can ensure *explainability of the AI system's decision-making process* (cf. III.B.). However, this last measure presupposes

[101] Min Kyung Lee, 'Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management' (2018) 5(1) Big Data & Society 1-16.

[102] Alon-Barkat and Busuioc, 'Human-AI Interactions in Public Sector Decision-Making', pp. 157-64.

[103] Ibid., p. 165.

[104] Liel and Zalmanson, 'Turning Off Your Better Judgment', p. 33.

[105] Cf. already Mosier and Skitka, 'Human Decision Makers and Automated Decision Aids', pp. 203-4.

[106] Ibid., p. 205 where the "extent to which organisations encourage the use of automated systems" is listed as one determining factor.

first, that the AI system is not a complete "black box"[107] and second, that the explanation can be done understandably not only for experts, e.g., the AI's developers, but also for human agents who may not have specific technical knowledge[108]. Effective information and explainability may require some additional *training* of the human agent.

When analysing these non-legal studies, however, it is essential to keep in mind the *legal* context in which the analysis is conducted. For instance, the findings of the studies may suggest that, in certain scenarios, erroneous Human-AI-Decisions could potentially be avoided by deliberately *misleading* humans: To mitigate Algorithm Aversion, e.g., in the context of "human" decisions, it may appear helpful to tell humans that the AI advice stems from another human. If the system is a "black box", humans may get fake explanations, which could foster their trust. Furthermore, to change humans' attitude towards an AI system in a more subtle way, previous mistakes could be concealed from them, or the AI system could be set up in a way that makes it appear more "human" (cf. III.B.). However, principals are generally not required or even allowed to take such measures: The "possibility" of a measure also needs to be assessed in light of the legal framework. For instance, to the extent that the algorithmic nature of advice must be made transparent,[109] hiding the source of advice does not constitute a *legally* possible risk mitigator.

IV. "Reasonable" Measures: Limits of Risk Mitigation in Human-AI-Decisions

Principals are not obliged to take all "possible" measures. Their duties of care only require them to do what is "reasonable". As mentioned above, to set this standard, the BGH, for example, considers *inter alia* "economic effects".[110] The precise role of economic aspects in liability law is a matter of debate beyond the scope of this

[107] Cf. European Commission, 'White Paper on Artificial Intelligence – A European approach to excellence and trust', 19 February 2020, COM(2020) 65 final, p. 12, where "opacity ('black box-effect')" is cited as one of the "specific characteristics of many AI technologies".

[108] Cf. Mosier and Skitka, 'Human Decision Makers and Automated Decision Aids', p. 208; cf., also, Fraunhofer IAIS, *Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz*, p. 67, where the necessity to distinguish between different groups of users is highlighted.

[109] Cf. e.g., Art. 13(2)(f) of the Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation); Art. 50 AI Act.

[110] Cf. notes 46 and 48 above.

paper.[111] All in all, it seems convincing to take them into account and even use them as a starting point, while bearing in mind they may need to be complemented, e.g., by social aspects.[112]

From an economic point of view, liability law should generally "reduce the sum of the costs of accidents and the cost of avoiding accidents".[113] Therefore, in principle, a safety measure should be taken if its costs are lower than the potential damage costs it prevents.[114] According to the so-called "Learned Hand formula", proposed in the U.S. by Judge Learned Hand, a person who does not take a risk prevention measure, acts negligently if "$B < (P \times L)$", where "B" refers to the burden of taking the measure, "P" to the probability of damage if the measure is not taken and "L" to the magnitude of such damage.[115] The usefulness of this formula is controversial.[116] However, it appears that courts, while refraining from explicitly citing "Learned Hand", at least tend to base their decisions implicitly on such cost-utility tests.[117]

Regarding Human-AI-Decisions, economic considerations allow to draw at least the following conclusions which could provide additional guidelines to define the

[111] For an overview on the discussion on economic analysis of tort law in Germany, cf. Jochen Taupitz, 'Ökonomische Analyse und Haftungsrecht – Eine Zwischenbilanz' (1996) 196(1/2) AcP 114-67; Gerhard Wagner, 'Vor § 823 BGB', in Jürgen Säcker et al. (eds.), *Münchener Kommentar zum Bürgerlichen Gesetzbuch*, 9th edn., 13 vols. (Munich: C.H. Beck, 2024), vol. VII, paras 66-76; Hans-Bernd Schäfer and Claus Ott, *Lehrbuch der ökonomischen Analyse des Zivilrechts*, 6th edn. (Berlin et al.: Springer Gabler, 2020), pp. 165-71.

[112] The need to complement the economic analysis is highlighted by Beckers and Teubner, *Three Liability Regimes for Artificial Intelligence*, pp. 16-7; Thomas M. J. Möllers, Juristische Methodenlehre, 5th edn. (Munich: C.H. Beck, 2023), p. 215-6; cf. equally the idea of a "socioeconomic analysis of law" mentioned by Josef Essers and Eike Schmidt, *Schuldrecht*, vol. I/1, 8th edn. (Heidelberg: C.F. Müller, 1995), p. 39; Taupitz, 'Ökonomische Analyse und Haftungsrecht', p. 126; cf., also, Martin Sommer, *Haftung für autonome Systeme* (Baden-Baden: Nomos, 2020), p. 270-2 (and passim) who convincingly suggests a "normativised" "risk-utility test".

[113] Gudio Calabresi, *The Costs of Accidents* (Cumberland: Yale University Press, 1970), p. 26; cf. Schäfer and Ott, *Lehrbuch der ökonomischen Analyse des Zivilrechts*, p. 170; Wagner, 'Vor § 823 BGB', para 65.

[114] Cf. Wagner, 'Vor § 823 BGB', para 66.

[115] *United States v Carroll Towing Co.*, 159 F.2d 169 (1947). The formula is named after the author of the opinion, Judge *Learned Hand*, cf. Schäfer and Ott, *Lehrbuch der ökonomischen Analyse des Zivilrechts*, pp. 202-3.

[116] Cf. Schäfer and Ott, *Lehrbuch der ökonomischen Analyse des Zivilrechts*, pp. 204-10.

[117] Cf. the analysis in Schäfer and Ott, *Lehrbuch der ökonomischen Analyse des Zivilrechts*, 204-6 and Gerhard Wagner, '§ 823 BGB', para 531. The formula may be supplemented by other (economic) considerations, especially when the victim can equally take measures to avoid accidents, cf. Schäfer and Ott, *Lehrbuch der ökonomischen Analyse des Zivilrechts*, pp. 279-84 with further references, where the formula is complemented by the idea of the "cheapest cost avoider".

University of Vienna Law Review, Vol. 9 No 3 (2025), pp. 35-63 https://doi.org/10.25365/vlr-2025-9-3-35.

boundaries of duties of care. First, a distinction must be made between the *enterprises*. The size of "*B*", the burden of the measure, depends not only on the amount of money to be paid but also on the paying organisation. For large and experienced enterprises in good financial standing, a safety measure is usually less of a burden than for small and medium-sized enterprises (SME) or start-ups with limited financial funds.[118] When determining negligence*, individual* deficiencies are not considered.[119] However, regarding the different "economic effects", one should distinguish between different *groups* of enterprises[120] and e.g., make higher demands on wealthy market leaders.[121]

Second, the *role of human intervention in the specific Human-AI cooperation* needs to be considered, as it determines "*P*", the probability of damage if the measure is not taken. In case of a Human-AI-Decision, the probability of damage depends on the skills of both the AI system and the human, and the extent to which the *principal*'s measures can reduce this probability depends on the interplay between these skills: When the AI system and the human make their individual assessments in *distinctive* ways, e.g., because they consider different aspects of the situation, a lack of human control (Automation Complacency, Automation Bias) or an excessive human control (Algorithm Aversion) is more likely to cause damage. In contrast, if they make their decisions *similarly*, "*P*" will not be reduced considerably by assuring that the human correctly exercises his or her own judgment, e.g., by providing more explainability.

Third, one must assess the *specific impact* of the Human-AI-Decision. The AI system could e.g., suggest a translation the human needs to accept or modify. "*L*", the magnitude of the damage is obviously higher when the translation concerns the use of a medical drug as in case of a Christmas card for a business partner.[122]

---

[118] The distinction is in line with the idea of reducing "secondary costs": A loss is generally less burdensome if the loss is spread among lots of entities (Loss Spreading Method) or if it concerns an entity which does not suffer a lot from the loss because it is in good financial standing (Deep Pocket Method), cf. Calabresi, *The Costs of Accidents*, pp. 39-45; Taupitz, 'Ökonomische Analyse und Haftungsrecht', pp. 140-1.

[119] Cf. BGH, 16 June 2009, VI ZR 107/08, BGHZ 181, 253-68, para 28 (juris); *Nettleship v Weston* 2 Q.B. 691 (1971) regarding English law; Schäfer and Ott, *Lehrbuch der ökonomischen Analyse des Zivilrechts*, p. 208; Wagner, '§ 823 BGB', paras 38-9.

[120] "Verkehrskreise", cf. BGH, 2 October 2012, VI ZR 311/11, BGHZ 195, 30-42, para 7 (juris); Wagner, '§ 823 BGB', para 40.

[121] Cf. Sommer, *Haftung für autonome Systeme*, pp. 238-42.

[122] Cf. the similar example by Fraunhofer IAIS, *Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz*, p. 12.

Finally, when weighing up the costs and benefits of a measure, it is important to remember that a measure that prevents one risk may at the same time create another risk. One example is training which increases *experience*: Experience may on the one hand lead to a better "subjective understanding"[123] and therefore improve humans' capacity to re-evaluate the AI advice (cf. III.A.1.). On the other hand, if the system works well, experience may also lead to "routine use"[124] and encourage Automation Complacency and Automation Bias (cf. III.C.). If the human sees the system err – which comes with more experience – this may foster Algorithm Aversion. At the same time, seeing it making mistakes in specific situations could lead to a better understanding and therefore reduce such Aversion (cf. III.B.). Resolving such trade-offs is challenging for both enterprises and courts.[125] The multidisciplinary approach will make it easier to define the duties of care, but not completely eliminate the difficulties.

A. Principals' Duties of Care Under the Proposal for an AI Liability Directive

In 2022, the European Commission presented a Proposal for a Directive on adapting non-contractual civil liability rules to artificial intelligence (AILD Proposal).[126] However, in February 2025, the Commission announced its withdrawal, stating that an agreement was not foreseeable.[127] The Commission will now "assess whether another proposal should be tabled or another type of approach should be chosen".[128] As will be shown in the following, the application to Human-AI-Decisions is one point where the approach of the AILD Proposal should indeed be reconsidered.

---

[123] Cf. note 96 above.

[124] Cf. notes 102 and 103 above.

[125] Cf. Schäfer and Ott, *Lehrbuch der ökonomischen Analyse des Zivilrechts*, p. 419 regarding general product liability law.

[126] European Commission, 'Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive)', 28 September 2022, COM(2022) 496 final (AILD Proposal). On the same day, the European Commission presented a 'Proposal for a Directive of the European Parliament and of the Council on liability for defective products', COM(2022) 495 final (PLD Proposal). Unlike the proposed AILD, the new PLD was adopted in 2024, cf. Directive (EU) 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.

[127] European Commission, 'Annexes to the Communication from the Commission to the European parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Commission work programme 2025', 11 February 2025, COM(2025) 45 final, Annex IV, p. 26.

[128] Ibid, p. 26.

The Proposal does not contain a new form of AI liability but seeks to harmonise *fault-based liability* by laying down common rules on the disclosure of evidence and the burden of proof.[129] According to the Recitals and the explanations, the AILD Proposal does not want to touch on the definition of "fundamental concepts" like "fault".[130] However, Art. 4(1)(a) AILD Proposal presupposes that in national tort law "fault" could consist "in the non-compliance with a duty of care laid down in Union or national law directly intended to protect against the damage that occurred".[131] The AILD Proposal does not lay out duties of care itself.[132] It frequently refers to the definitions and obligations set out in the AI Act (cf. Art. 2, Art. 4(2) and (3) AILD Proposal).

Under the AILD Proposal, non-compliance with a duty of care can be both an *effect* and a *prerequisite* of a (rebuttable) presumption in favour of the victim: First, where a defendant fails to comply with a court order to disclose or preserve evidence, the court shall presume the defendant's breach of a duty of care (Art. 3(5) AILD Proposal – *effect*). Second, when the fault of the defendant is established and consists in the non-compliance with a duty of care laid down in Union or national law directly intended to protect against the damage that occurred, courts shall – under certain other conditions – presume a causal link between the fault and the output of the AI system (Art. 4(1) AILD Proposal – *prerequisite*). However, it is questionable whether the AILD Proposal applies to the Human-AI-Decisions analysed in this paper: Recital 15 suggests that it does not cover "liability claims when the damage is caused by a human assessment followed by a human act or omission, while the AI system only provided information or advice which was taken into account by the relevant human actor". This restriction has been criticised[133] and the critic's arguments are convincing: As seen above, the impact of the human "in the loop" can be illusory.[134] The AILD Proposal's assumption that in such cases "it is possible to trace back the

[129] Cf. Art. 1 AILD Proposal. However, according to Art. 5(1) and (2), five years after the end of the transition period, the Commission must present a report that should in particular "evaluate the appropriateness of no-fault liability rules for claims against the operators of certain AI systems".

[130] Cf. Explanatory Memorandum of the AILD Proposal, p. 11; Recital 10 AILD Proposal.

[131] Cf. De Bruyne et al., 'The European Commission's approach to extra-contractual liability and AI', p. 8: "the reliance on this concept [duty of care] in EU legislation is rather surprising".

[132] Art. 2(9) AILD Proposal defines a "duty of care" as "a required standard of conduct, set by national or Union law, in order to avoid damage to legal interests recognised at national or Union law level, including life, physical integrity, property and the protection of fundamental rights".

[133] Philipp Hacker, 'The European AI liability directives – Critique of a half-hearted approach and lessons for the future' (2023) 51 Computer Law & Security Review 105871, pp. 13-4.

[134] Ibid., p. 13.

University of Vienna Law Review, Vol. 9 No 3 (2025), pp. 35-63 https://doi.org/10.25365/vlr-2025-9-3-35.

damage to a human act or omission, as the AI system output is not interposed between the human act or omission and the damage" may be true but misses the actual problem of the victim. Proving the "factual" causation[135] between a human act or omission – which could also consist in the simple activation of the AI system – and a damage is usually possible.[136] The real difficulty consists in establishing whether the human act or omission breached a duty of care and whether the damage would not have occurred if this duty had been respected[137]. This difficulty also exists for victims of Human-AI-Decisions: As shown above, there are many factors to be considered when cooperating with AI systems and when organising such Human-AI cooperation. The victim's typical lack of insight into first, the AI system, second the Human-AI cooperation and third, the specific organisation this cooperation is integrated into, presents significant obstacles.[138] The wording of Art. 3 AILD Proposal would allow to extent the *disclosing obligations* and the *presumption of non-compliance* in case of a violation to these victims: It includes any "high-risk AI system that is suspected of having caused damage" (Art. 3(1) AILD Proposal) and any "claim for damages" (Art. 3(2) AILD Proposal).[139] There is no requirement that the claim must be based on a faulty human behaviour *preceding* the AI output. In contrast, the *presumption of causality* in the case of fault contained in Art. 4 AILD Proposal seems to be tailored to fully automated AI systems: It only allows to presume the "causal link between the fault of the defendant and the output produced by the AI system". Taken literally, the presumption does not provide any relief if the faulty human behaviour *succeeds* the AI output.[140] Consequently, it would not cover claims against the human agent directly cooperating with the AI system. However, the wording of Art. 4 AILD Proposal would already allow applying the presumption to claims against the

[135] Cf. Brüggemeier, *Haftungsrecht: Struktur, Prinzipien, Schutzbereich,* p. 28.

[136] Cf. Tianyu Yuan, 'Lernende Roboter und Fahrlässigkeitsdelikt' (2018) 9(4) RW 477-504, p. 493.

[137] This latter aspect is part of the "legal" causation, cf. Brüggemeier, *Haftungsrecht: Struktur, Prinzipien, Schutzbereich,* p. 28.

[138] Cf. Hacker, 'The European AI liability directives', pp. 13-4: "it must be questioned whether it is really equally difficult to prove causality and non-compliance in cases of human intervention with and without AI involvement"; cf. equally De Bruyne et al., 'The European Commission's approach to extra-contractual liability and AI', p. 18 highlighting that – with regards to the Art. 8 PLD Proposal – "the complexity at stake is rather of an *organisational* nature and/or relates to the information asymmetry that the consumer endures concerning the apportion of responsibilities between the various actors at stake".

[139] Cf. Hacker, 'The European AI liability directives', p. 14: "Hence, the presumption of non-compliance contained in Article 3(5) AILD Proposal should apply equally if a human agent took, or failed to take, the final decision leading to the damage caused by the AI output".

[140] Cf. ibid., p. 23.

University of Vienna Law Review, Vol. 9 No 3 (2025), pp. 35-63 https://doi.org/10.25365/vlr-2025-9-3-35.

principals of these human agents: *Their* faulty behaviour, e.g., a violation of a duty to inform the human, usually *precedes* the AI output.[141] Nevertheless, a potential new proposal or other type of approach to AI Liabilitlyshould at least reconsider this limitation.[142]

B. Summary and Perspectives

Enterprises increasingly make use of Human-AI-Decisions. When such decisions result in damage the question of the enterprise's liability arises. This paper has focused on enterprises' non-contractual and *fault-based* liability which generally requires a violation of a *duty of care* by the principal. When it comes to Human-AI-Decisions, it is challenging for courts to define these duties of care. The paper has shown how findings from other disciplines can help judges to determine "possible" and "reasonable" measures principals must take to prevent specific risks of Human-AI-Decision-making. There are three main sources of such risks: Automation Complacency and Automation Bias on the one hand and Algorithm Aversion on the other hand. Scientists have conducted numerous studies on these phenomena and have identified mitigating factors. Their findings are part of the "state of scientific and technical knowledge" and need to be considered when setting legal safety standards. Nonetheless, principals do not need to take every "possible" measure. To determine if a measure can be reasonably expected, knowledge from other disciplines, in particular from economics, is again of considerable value.

At the same time, this paper has shown that some challenges remain, especially when it comes to resolving trade-offs associated with a particular measure. The now abandoned AILD Proposal would have provided little relief in this regard. Thus, the question as to whether liability for AI systems should not be tightened *de lege ferenda* remains pressing.[143]

---

[141] Whether the Recitals can limit the scope of application of a directive cannot be discussed here; for an analysis of the significance of recitals in the methodology of EU law, cf. Tobias Gumpp, 'Stellenwert der Erwägungsgründe in der Methodenlehre des Unionsrechts' (2022) 8(4) ZfPW 446-76.

[142] Cf. Hacker, 'The European AI liability directives', p. 23: "What would be needed is a presumption that the act/omission of the user [..] caused the damage".

[143] Cf. on this issue e.g., Gerhard Wagner, 'Verantwortlichkeit im Zeichen digitaler Techniken' (2020) 71 VersR 717-41, pp. 734-41; Zech, 'Entscheidungen digitaler autonomer Systeme', pp. A 87-110; Beckers and Teubner, *Three Liability Regimes for Artificial Intelligence*, pp. 45-166; Benedetta Cappiello, *AI-systems and non-contractual liability* (Torino, Giappichelli: 2022), pp. 45-96; Mayrhofer, *Außervertragliche Haftung für fremde Autonomie*, pp. 370-443 and 'Product liability in the age of AI — Proposal for a "two track" solution' (2024) 33(1) *Revista Electrónica de Direito* 105-127; cf., also, Philipp Hacker, *Proposal for a directive on adapting non-contractual civil liability rules to artificial intelligence, Complementary impact assessment* (2025), available at

V. Bibliography

Alon-Barkat, Saar and Busuioc, Madalina, 'Human-AI Interactions in Public Sector Decision-Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice' (2023) 33(1) *Journal of Public Administration Research and Theory* 153–69

Ball, Matthew and Callaghan, Vic, 'Explorations of Autonomy', in *2012 Proceedings of the 8th International Conference on Intelligent Environments* (Guantajo: IEE Xplore, 2012) 114-21

Bauer, Kevin, et al., 'Die KI braucht bei der Bankberatung immer noch menschliche Hilfe', *Börsen-Zeitung*, available at https://www.boersen-zeitung.de/kapitalmarktforschung/in-der-bankberatung-braucht-die-ki-menschliche-hilfe-90edbb42-86a4-11ed-a311-f90ecc32c8e4 (last accessed 25 April 2025)

Beckers, Anna and Teubner, Gunther, *Three Liability Regimes for Artificial Intelligence* (Oxford: Hart, 2022)

Bomhard, David and Merkle, Marieke, 'Europäische KI-Verordnung – Der aktuelle Kommissionsentwurf und praktische Auswirkungen' (2021) 1(6) *Recht Digital (RDi)* 276-282

Börütecene, Ahmet and Löwgren, Jonas, 'Designing Human-Automation Collaboration for Predictive Maintenance', in *Companion Publication of the 2020 ACM Designing Interactive Systems Conference* (New York: Association for Computing Machinery, 2020) 251-6

Brüggemeier, Gert *Haftungsrecht: Struktur, Prinzipien, Schutzbereich* (Berlin et al.: Springer, 2006)

Calabresi, Gudio, *The Costs of Accidents* (Cumberland: Yale University Press, 1970)

Cappiello, Benedetta, *AI-systems and non-contractual liability* (Torino, Giappichelli: 2022)

Choi, Benard C.K. and Pak, Anita W.P., 'Multidisciplinarity, interdisciplinarity, and transdisciplinarity in health research, services, education and policy: 1. Definitions, objectives, and evidence of effectiveness' (2006) 29(6) *Clinical and Investigative Medicine* 351-64

Cummings, Mary L., 'Automation Bias in Intelligent Time Critical Decision Support Systems' (2004) *AIAA 1st Intelligent Systems Technical Conference, 20-22*

https://www.europarl.europa.eu/RegData/etudes/STUD/2024/762861/EPRS_STU(2024)762861_EN.pdf (last accessed 25 April 2025).

September 2004, Chicago, IL, American Institute of Aeronautics and Astronautics 1-6

Datenethikkommission der Bundesregierung, *Gutachten* (2019), available at https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.html (last accessed 25 April 2025)

De Bruyne, Jan, et al., 'The European Commission's approach to extra-contractual liability and AI – An evaluation of the AI liability directive and the revised product liability directive' (2023) 51 *Computer Law & Security Review* 105894

Deutsches Institut für Normung and Deutsche Kommission Elektrotechnik, Elektronik, *Deutsche Normungsroadmap Künstliche Intelligenz*, 2nd edn. (2022), available at https://www.din.de/de/forschung-und-innovation/themen/kuenstliche-intelligenz/fahrplan-festlegen (last accessed 25 April 2025)

Dietvorst, Berkeley J., et al., 'Algorithm aversion: People erroneously avoid algorithms after seeing them err' (2015) 144(1) *Journal of Experimental Psychology: General* 114-26

Essers, Josef and Schmidt, Eike, *Schuldrecht*, vol. I/1, 8th edn. (Heidelberg: C.F. Müller, 1995)

Expert Group on Liability and New Technologies, 'Liability for Artificial Intelligence and Other Emerging Digital Technologies' (2019), available at https://op.europa.eu/en/publication-detail/-/publication/1c5e30be-1197-11ea-8c1f-01aa75ed71a1/language-en (last accessed 25 April 2025)

Eykholt, Kevin et al., 'Robust Physical-World Attacks on Deep Learning Visual Classification', in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Salt Lake City: IEEE, 2018) 1625-34

Fawcett, Jacqueline, 'Thoughts About Multidisciplinary, Interdisciplinary, and Transdisciplinary Research' (2013) 26(4) *Nursing Science Quarterly* 376–9

Fraunhofer IAIS, *Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz* (2021), available at https://www.iais.fraunhofer.de/de/publikationen/studien/2021/ki-pruefkatalog.html (last accessed 25 April 2025)

Fügener, Andreas et al., 'Cognitive Challenges in Human–Artificial Intelligence Collaboration: Investigating the Path Toward Productive Delegation' (2022) 33(2) *Information Systems Research* 678–96

Goddard, Kate, et al., 'Automation bias: a systematic review of frequency, effect mediators, and mitigators' (2012) 19(1) *Journal of the American Medical Informatics Association* 121-7

Grace, Katja, et al., 'Viewpoint: When Will AI Exceed Human Performance? Evidence from AI Experts' (2018) 62 *Journal of Artificial Intelligence Research* 729-54

Gsenger, Rita and Strle, Tomam, 'Trust, Automation Bias and Aversion: Algorithmic Decision-Making in the Context of Credit Scoring' (2021) 19(4) *Interdisciplinary Description of Complex Systems* 542-60

Gumpp, Tobias, 'Stellenwert der Erwägungsgründe in der Methodenlehre des Unionsrechts' (2022) 8(4) *Zeitschrift für die gesamte Privatrechtswissenschaft (ZfPW)* 446-76

Hacker, Philipp, *Proposal for a directive on adapting non-contractual civil liability rules to artificial intelligence, Complementary impact assessment* (2025), available at https://www.europarl.europa.eu/RegData/etudes/STUD/2024/762861/EPRS_STU(2024)762861_EN.pdf (last accessed 25 April 2025)

Hacker, Philipp, 'The European AI liability directives – Critique of a half-hearted approach and lessons for the future' (2023) 51 *Computer Law & Security Review* 105871

Hacker, Philipp, 'Verhaltens- und Wissenszurechnung beim Einsatz von Künstlicher Intelligenz' (2018) 9(3) *Rechtswissenschaft (RW)* 243-88

High-Level Expert Group on Artificial Intelligence, 'Ethics Guidelines for Trustworthy AI' (2019), available at https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (last accessed 25 April 2025)

Hilgendorf, Eric, 'Bedingungen gelingender Interdisziplinarität' (2010) 65(19) *Juristenzeitung (JZ)* 913-22

Hofmann, Franz, 'Disziplinarität, Intradisziplinarität und Interdisziplinarität am Beispiel der Grundsätze „mittelbarer Verantwortlichkeit"' (2018) 73(15-6) *JuristenZeitung (JZ)* 746-54

Jussupow, Ekaterina et al., 'Augmenting Medical Diagnosis Decisions? An Investigation into Physicians' Decision-Making Process with Artificial Intelligence' (2021) 32(3) *Information System Research* 713-35

Kahneman, Daniel et al., 'NOISE' (2016) *40 Harvard Business Review* 38-46

Kaminski, Andreas, 'Gründe geben. Maschinelles Lernen als Problem der Moralfähigkeit von Entscheidungen', in Klaus Wiegerling et al. (eds.), *Datafizierung und Big Data* (Wiesbaden: Springer VS, 2020) 151-74

Koziol, Helmut (ed.), *Basic Questions of Tort Law from a Comparative Perspective* (Vienna: Sramek, 2015)

Koziol, Helmut (ed.), *Basic Questions of Tort Law from a Germanic Perspective* (Vienna: Sramek, 2012)

Koziol, Helmut (ed.), *Comparative Stimulations for Developing Tort Law* (Vienna: Sramek, 2015)

Koziol, Helmut, 'Comparative Conclusions' in Helmut Koziol (ed.), *Basic Questions of Tort Law from a Comparative Perspective* (Vienna: Sramek, 2015) 685-838

Koziol, Helmut, 'Concluding Remarks', in Helmut Koziol (ed.*), Comparative Stimulations for Developing Tort Law* (Vienna: Sramek, 2015) 182-95

Koziol, Helmut, et al., 'Liability for Agents and Agents' Liability', in Helmut Koziol (ed.), *Comparative Stimulations for Developing Tort Law* (Vienna: Sramek, 2015) 182-95

Larson, Erik J., *The myth of artificial intelligence* (Cambridge et al.: Harvard University Press, 2021)

Lee, Min Kyung, 'Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management' (2018) 5(1) *Big Data & Society* 1-16

Liel, Yotam and Zalmanson, Lior, 'Turning Off Your Better Judgment – Conformity to Algorithmic Recommendations' (Working paper) (2022), available at https://www.researchgate.net/publication/366412145_Turning_Off_Your_Better_Judgment_-Conformity_to_Algorithmic_Recommendations (last accessed 25 April 2025)

Mayrhofer, Ann-Kristin, 'Product liability in the age of AI — Proposal for a "two track" solution' (2024) 33(1) *Revista Electrónica de Direito* 105-127

Mayrhofer, Ann-Kristin, *Außervertragliche Haftung für fremde Autonomie* (Tübingen: Mohr Siebeck, 2023)

Milner, Greg , 'Death by GPS: Are Satnavs changing our brains?', *The Guardian*, available at https://www.theguardian.com/technology/2016/jun/25/gps-horror-stories-driving-satnav-greg-milner (last accessed 25 April 2025)

Möllers, Thomas M. J., *Juristische Methodenlehre*, 5th edn. (Munich: C.H. Beck, 2023)

Mori, Mashiro, 'The Uncanny Valley' (2012) 19(2) *IEEE Robotics & Automation Magazine* 98–100

Mosier, Kathleen L. and Skitka, Linda J., 'Human Decision Makers and Automated Decision Aids: Made for Each Other?', in Raja Parasuraman and Mustapha Mouloua (eds.), *Automation and human performance: Theory and application (*Mahwah: Lawrence Erlbaum, 1996) 201–20

Parasuraman, Raja and Manzey, Dietrich H., 'Complacency and Bias in Human Use of Automation: An Attentional Integration' (2010) 53(3) *Human Factors* 381-410

Schäfer, Hans-Bernd and Ott, Claus, *Lehrbuch der ökonomischen Analyse des Zivilrechts*, 6th edn. (Berlin et al.: Springer Gabler, 2020)

Sommer, Martin, *Haftung für autonome Systeme* (Baden-Baden: Nomos, 2020)

Stürner, Rolf, 'Die Zivilrechtswissenschaft und ihre Methodik – zu rechtsanwendungsbezogen und zu wenig grundlagenorientiert?' (2014) 214(1-2) *Archiv für die civilistische Praxis (AcP)* 7-54

Susanne Horner and Markus Kaulartz, 'Haftung 4.0' (2016) 32(1) *Computer und Recht (CR)* 7-14

Taupitz, Jochen, 'Ökonomische Analyse und Haftungsrecht – Eine Zwischenbilanz' (1996) 196(1/2) *Archiv für die civilistische Praxis (AcP)* 114-67

Teodorescu, Mike H. M., et al., 'Failures of Fairness in Automation Require a Deeper Understanding of Human-ML Augmentation' (Minneapolis: University of Minnesota 2021) 45(3) *MIS Quarterly* 1483-500

Wagner, Gerhard, 'Roboter als Haftungssubjekte? Konturen eines Haftungsrechts für autonome Systeme', in Florian Faust and Hans-Bernd Schäfer (eds.), *Zivilrechtliche und rechtsökonomische Probleme des Internet und der künstlichen Intelligenz* (Tübingen: Mohr Siebeck, 2019) 1-39

Wagner, Gerhard, 'Verantwortlichkeit im Zeichen digitaler Techniken' (2020) 71 *Versicherungsrecht (VersR)* 717-41

Wagner, Gerhard, 'Vor § 823 BGB' and '§ 823 BGB', in Jürgen Säcker et al. (eds.), *Münchener Kommentar zum Bürgerlichen Gesetzbuch*, 9th edn., 13 vols. (Munich: C.H. Beck, 2024), vol. VII

Yeomans, Michael, et al., 'Making sense of recommendations' (2019) 32(4) *Journal of Behavioral Decision Making* 403-14

Yuan, Tianyu, 'Lernende Roboter und Fahrlässigkeitsdelikt' (2018) 9(4) *Rechtswissenschaft (RW)* 477-504

Zech, Herbert, 'Entscheidungen digitaler autonomer Systeme: Empfehlen sich Regelungen zu Verantwortung und Haftung?', in Ständige Deputation des Deutschen Juristentages (ed.), *Verhandlungen des 73. Deutschen Juristentages* (Munich: C.H. Beck, 2020) A 1-112