AI for decision support:
What are possible futures, social impacts, regulatory options, ethical conundrums and agency constellations?

KI zur Entscheidungsunterstützung:
Was sind mögliche Zukünfte, soziale Auswirkungen, regulatorische Optionen, ethische Fragen und Akteurskonstellationen?

Edited by D. Schneider and K. Weber
INTRODUCTION

AI-based decision support systems and society: An opening statement

Diana Schneider¹, Karsten Weber*²

Abstract • Although artificial intelligence (AI) and automated decision-making systems have been around for some time, they have only recently gained in importance as they are now actually being used and are no longer just the subject of research. AI to support decision-making is thus affecting ever larger parts of society, creating technical, but above all ethical, legal, and societal challenges, as decisions can now be made by machines that were previously the responsibility of humans. This introduction provides an overview of attempts to regulate AI and addresses key challenges that arise when integrating AI systems into human decision-making. The Special topic brings together research articles that present societal challenges, ethical issues, stakeholders, and possible futures of AI use for decision support in healthcare, the legal system, and border control.


Keywords • artificial intelligence (AI), decision support, socio-technical systems, regulation, social impacts.

This article is part of the Special topic "AI for decision support: What are possible futures, social impacts, regulatory options, ethical conundrums and agency constellations?*, edited by D. Schneider and K. Weber. https://doi.org/10.14512/tatup.33.1.08

Introduction

In recent years the use of artificial intelligence (AI) systems to support decision-making has become established in various areas of application and has therefore also gained societal significance, as more and more individuals are affected by AI systems in very different situations and contexts. In contrast to systems that decide certain aspects autonomously, decision support systems (DSS) are characterized by the fact that they are merely a decision-making aid for human users. By means of AI, for example, decisions can be prepared by analyzing large amounts of data or recognizing patterns in it. While this can increase the efficiency and accuracy of decisions, it could have a variety of serious and far-reaching effects on individuals, groups, institutions, associations, companies, and society as well as the natural environment.

Since the scope of the impacts and the number of parties affected is so vast, or at least appears to be so vast, extreme scenarios are all too often conjured up in which AI systems either subjugate humanity or solve all of humanity’s pressing problems, from climate change to combating pandemics, especially in public discussions about AI. The recent debate on large language models in general and ChatGPT in particular also follows this pattern, with proponents – to put it somewhat tongue-in-cheek – declaring the use of AI systems a panacea and opponents label-
It is currently impossible to predict what effects the proposed EU regulation will have on other countries and within the international discourse.

Attempts to regulate artificial intelligence

The research articles in this Special topic focus predominantly on outlining possible areas of application for (AI-based) decision support systems, identifying stakeholders and describing any potential impacts. Given the novelty of the subject, this is not only an important but also a difficult task. As a result, regulatory issues could only be dealt with marginally or not at all; therefore, at least a few comments on this should be made.

When a provisional agreement on the Artificial Intelligence Act (AIA) was reached on 9 December 2023 after lengthy negotiations in a trialogue between the European legislator, the European Parliament and the Council of the EU, this was heralded as a historic moment in the regulation of AI. The AIA takes a risk-based approach to regulation: While AI systems with no or only low risk are hardly regulated, special requirements apply to high-risk applications, e.g., specific transparency obligations and extensive requirements for data quality, documentation, and traceability (European Union 2023; European Commission 2021). Transparency is considered to be highly relevant in order to interpret AI-generated results and ensure appropriate use (European Commission 2021, reason 47, p. 30) – and thus ultimately contributes to the explainability of AI analyses and recommendations, so seems to be the assumption of EU lawmakers.

The measures proposed in the AIA are primarily aimed at preventing potential risks to the fundamental rights, health, or safety of EU citizens. The debate on the draft regulation is particularly essential for the discussion of AI-supported decision-making systems, as all the fields of application covered in this TATuP Special topic (jurisdiction, law enforcement, and medicine) in principle must be considered particularly sensitive areas. The AI use cases discussed in the research articles can therefore have a considerable impact on the lives of those affected – not only in the event of an error, but also in regular use.

Yet, the AIA was and is on no account the only attempt to regulate the use of AI (Butcher and Beridze 2019; Schift et al. 2022; Schmitt 2022; Ulnicane et al. 2021). An arbitrary and by no means comprehensive selection shows the range of types of actors and approaches to regulation: For instance, the OECD has formulated the Recommendation of the Council on Artificial Intelligence, the EU the Ethics Guidelines for Trustworthy AI and the Future of Life Institute the Asilomar AI Principles. These and many other documents appear to propose ethical guidelines and codes of ethics for regulation – at least that is what an initial review suggests. However, the binding force of ethical guidelines and codes of ethics is based on voluntary commitment; there is therefore no enforceability that only laws could offer. Moreover, Schift et al. (2022) emphasize that most of these documents offer little indication of how requirements, recommendations and/or claims they propose can be translated into actionable instructions for the practice of AI development and use, and instead remain at the rather abstract level of moral imperatives.
It can only be assumed that, in view of competing regulatory approaches, the AIA is rather not the last word on the regulation of AI, even more so as criticism of the AIA has not been long in coming. Furthermore, it is currently impossible to predict whether and what effects the proposed EU regulation will have on other countries and within the international discourse. How the application of AI will be regulated in, say, ten years’ time in the areas covered in this Special topic as well as in other areas is therefore difficult to predict today.

Human decisions and artificial intelligence

With regard to the question of how AI systems can be specifically integrated into human decision-making, mainly theoretical considerations and only a few empirical studies exist. Many of the following considerations originate from the medical context, as the impact of AI systems on human decision-making has long been the subject of intensive research, particularly in the healthcare professions. For example, Braun et al. (2020) outlined various modes of interaction for the healthcare sector (e.g., the integrative AI-DSS, which can independently request and collect patient data, or the fully automated AI-DSS, which does not require the involvement of professionals); however, most considerations primarily assume direct, essentially bilateral interaction between the professional and the AI system, i.e., a conventional AI-DSS. Simultaneously, there is widespread agreement that the integration of AI systems into the professional decision-making process—regardless of the respective mode of interaction—will have an impact on established work relations, e.g., on the relationship between professionals and patients or employees and employers (Schneider et al. 2022b). On the one hand, the use of DSS is expected to reduce the workload and the potential time saved is associated with a more empathetic approach to patients (Topol 2019) – expectations that appear questionable given the increasing costs of purchasing and maintaining technology and educating personnel as well as labor shortages. On the other hand, there are concerns that computer paternalism could undermine the essential relationship of trust between healthcare professionals and patients (Cartolovni et al. 2022; Heyen and Salloch 2021) – or, to put it in more general terms, between professionals and clients. Studies already indicate that time and again automation bias occurs (Sujan et al. 2019), i.e., recommendations from AI-based decision-making systems are adopted without question. In view of the frequent lack of data literacy, this poses an enormous challenge, as the AI-generated recommendations can only be used responsibly if they can be correctly understood and interpreted by the users. The perception of AI systems as a second opinion could also raise further ethical questions regarding responsibility (Kempt and Nagel 2022). How AI-based systems could be meaningfully incorporated into shared decision-making processes (e.g., between patients or clients and professionals) also appears to be largely unresolved.

As the use of AI-based systems results in a stronger focus on data and the information, patterns, or meta-information it contains, other forms of professional knowledge could become jeopardized. Particularly in areas of application in which human experience, intuition, tactile or implicit knowledge, but also interpersonal relationships are highly valued (e.g., in the social and healthcare sector as well as in case of judicial or administrative decisions), an inappropriate focus on dataism is a concern and has been strongly criticized in some cases (Pedersen 2019; Devlieghere et al. 2022; Webb 2003). Various papers have pointed out that the data sets used for AI-based systems are fragmented (Tucker 2023), may contain deliberate omissions (Schneider 2022), or that administrative data sets are unsuitable for assessing professional issues (Gillingham 2015, 2020). Besides the fact that most training datasets have a strong bias and are poorly representative in terms of ethnic origin and gender, for example, there is also the problem that particularly vulnerable and/or stigmatized groups of people are often underrepresented.

However, analyzing large data sets opens up the possibility of contributing to evidence-based practice. But this requires that the algorithms underlying the AI-based recommendations are not exclusively pattern-based, but also incorporate concepts and theories from current research and knowledge—otherwise it would be almost impossible to make valid statements (Schneider et al. 2022a).

If AI systems suggest decisions and users routinely adopt them, it is not the technology that has changed but the way it is used.

These short comments can only highlight a few aspects regarding the use of AI systems for decision support. For instance, the differentiation between systems for automatic decision-making (human-out-of-the-loop) and for decision support (human-in-the-loop, human-on-the-loop) should certainly be dealt with in much more detail, as different questions are raised depending on how AI systems are actually employed. It would also be worth to examine whether and how transitions from decision support systems to automatic decision-making systems might take place; this is less a technical issue than an organizational and practical one, because if AI systems suggest decisions but users routinely adopt them, it is not the technology that has changed but the way it is used. Such transitions in modes of use can in turn lead to far-reaching changes in the respective understanding of the profession and this in turn can again change modes of use.
(Schneider et al. 2022 b). In other words: When talking about AI, this must always be done in terms of a socio-technical system.

Contributions in this Special topic

The six contributions to this TATuP Special topic cover the use of AI systems in three different domains: healthcare, legal system, and law enforcement or, more precisely, border control. We decided to cluster thematically related research articles and to sort them in the clusters according to the alphabetical order of the names of the first authors – this seemed to us to be the best variant of an ultimately arbitrary arrangement. The first three research articles deal with the use of AI in the legal system, followed by a research article on AI systems used to identify illegal migrants at the border, and finally two research articles on the application of AI systems in medical contexts.

- Anna-Katharina Dhungel and Moreen Heine are investigating whether, how and who would benefit from the use of AI systems in the legal system. They will attempt to answer this question based on interviews with judges from Germany. On the one hand, this means a limitation, as the results are only meaningful in the context of the German legal system, yet on the other hand, comparable studies have so far mostly been carried out in countries, such as the United States, whose legal system is structured significantly differently to the German system. The article therefore fills a research gap.

- Brandon Long and Amitabha Palmer focus on a different target group, as they look at the question of whether AI systems could have an advantage for jurisdiction from the perspective of the users of the legal system in the United States. Their primary interest is whether AI systems could be used as cost-effective advisors, in particular for socially disadvantaged people, providing, for example, reliable information about the potential success of a lawsuit.

- With reference to the legal system, Giovana Lopes asks whether AI systems could be used to identify prejudices and partialities – bias for short – among judges, to alert them to the existence of bias, thereby inducing behavioral and attitudinal changes on the part of judges and thus ensuring fairer judgments.

- In their research article on the use of AI in the context of border controls, Daniel Minkin and Lou Therese Brandner explore the theoretical assumptions underlying the ‘iBorderCtrl’ system. They conclude that the theoretical foundations of the system are questionable and that the use of the system to recognize false statements made by people entering a country should therefore be called into question. This indicates that in the context of technology assessment not only technology, but also fundamental and/or theoretical assumptions that are incorporated into technology should be taken into account.

- Manuela Marquardt, Philipp Graf, Eva Jansen, Stefan Hillmann, and Jan-Niklas Voigt-Antons use a scenario-based interview study to investigate which requirements AI systems in the medical context must fulfill in terms of explainability so that their outputs are comprehensible for all stakeholders. Without situationally adapted and comprehensible explanations, building trust in AI systems would hardly be possible and their use would therefore be called into question.

- Finally, Jan C. Zoellick, Hans Drexler, and Konstantin Drexler address the use of AI systems for the detection of melanoma. Using three scenarios, they consider shifts in competences from physicians to the systems used, examine the need for regulation of the use of AI systems in diagnostic contexts and finally look at a conundrum that is often brought up but rarely clearly resolved: What should be done if humans and machines come to different conclusions? Whose conclusion should take priority?

Conclusion

As already indicated, the contributions in this Special topic cannot address all aspects of AI-based decision-making support. Regarding technology assessment they do, however, provide examples of the topics and issues raised by the rapid and ubiquitous introduction of AI technologies. In the case of country-specific studies, the potential for drawing generalizations may be limited, but particularly from a technology assessment perspective, such specific studies cannot be dispensed with, as the effects of technology are determined to a considerable extent by the prevailing conditions. Discussions of specific topics usually differ not only from country to country, but also within a profession in a country where there are different and long-established strands of discourse with corresponding arguments and assumptions, e.g. in the field of the digitalization of social work in Germany (Waag and Rink 2023). Country comparisons and interdisciplinary studies can therefore help to make such discourses more comprehensible and transparent. The comparison of arguments and assumptions can also help to uncover blind spots in one’s own argumentation. A differentiated view that is considering country-specific characteristics and social conditions is also indispensable with regard to impact assessment and evaluation of technology, as otherwise there is a risk of not being able to move beyond thinking in terms of extreme scenarios. While the research articles in this TATuP Special topic refer to similar challenges and issues, they also illustrate the importance of detail and differentiation, despite the variety of subjects covered.

Funding - This work received no external funding.

Competing interests - Karsten Weber is a member of TATuP’s scientific advisory board. He was not involved in the editorial voting process for the article’s approval.

Acknowledgement

The Special topic editors would like to thank the authors and reviewers for the professional and most cooperative collaboration.
References


Schiff, Daniel; Laas, Kelly; Biddle, Justin; Borenstein, Jason (2022): Global AI ethics documents. What they reveal about motivations, practices, and policies. In: Kelly Laas, Michael Davis and Elizabeth Hildt (eds.): Codes of ethics and ethical guidelines. Cham: Springer International, pp. 121–143. https://doi.org/10.1007/978-3-030-86201-5_7


Schneider, Diana, Maier, Angelika; Cimiano, Philipp; Seelmeyer, Udo (2022a): Exploring opportunities and risks in decision support technologies for social workers. An empirical study in the field of disabled people’s services. In: Sozialer Fortschritt 71 (6–7), pp. 489–511. https://doi.org/10.3979/sfo.71.6-7/489


DIANA SCHNEIDER

is research associate at the Competence Center Emerging Technologies at the Fraunhofer Institute for Systems and Innovation Research ISI since 2021. She was a PhD candidate of the NRW Digital Society research program from 2018–2022. Her research focuses on innovations in social and healthcare systems, in particular their ethical and social implications.

PROF. DR. KARSTEN WEBER

is research professor for Technology Assessment and AI-based Mobility at OTH Regensburg since 2022 and has been working there as senior researcher since 2013. His research is focused on impacts of technology on individuals, groups, society, and environment, namely information and communication technology, particularly in healthcare, mobility, and energy.
Abstract • Despite substantial artificial intelligence (AI) research in various domains, limited attention has been given to its impact on the judiciary, and studies directly involving judges are rare. We address this gap by using 20 in-depth interviews to investigate German judges’ perspectives on AI. The exploratory study examines (1) the integration of AI in court proceedings by 2040, (2) the impact of increased use of AI on the role and independence of judges, and (3) whether AI decisions should supersede human judgments if they were superior to them. The findings reveal an expected trend toward further court digitalization and various AI use scenarios. Notably, opinions differ on the influence of AI on judicial independence and the precedence of machine decisions over human judgments. Overall, the judges surveyed hold diverse perspectives without a clear trend emerging, although a tendency toward a positive and less critical evaluation of AI in the judiciary is discernible.

Cui bono? Richterliche Entscheidungsfindung in Zeiten von KI: Eine qualitative Untersuchung zur Erwartungshaltung von Richter*innen in Deutschland


Keywords • artificial intelligence, algorithmic judges, user-centered studies, e-justice, expert interviews

This article is part of the Special topic “AI for decision support: What are possible futures, social impacts, regulatory options, ethical conundrums and agency constellations?,” edited by D. Schneider and K. Weber. https://doi.org/10.14512/tatup.33.1.08

Introduction

While significant strides have been made in the exploration of AI’s impact on various sectors, the judicial domain remains relatively understudied within this discourse. Existing research primarily centers on legal issues related to AI implementation in courts, public perceptions of algorithmic judges, and isolated technical case studies (Eidenmüller and Wagner 2021; Watson et al. 2023; Yalcin et al. 2023). Particularly, the deployment of risk assessment systems in criminal proceedings attracts scholarly attention, notably concerning issues of justice and discrimination, alongside the technical feasibility (Berk 2019; Dressel and Farid 2018; Završnik 2020). However, there is a conspicuous lack of studies that focus on the attitudes of judges towards AI. Publications that specifically address this audience are rare,
both in the context of Germany, and in comparison with other nations. Notably for the German-speaking region, Hartung et al. (2022) examined the future of digital justice, involving interviews with judges, and a publication by IBM compiles insights garnered from discussions with (vice-)presidents of various courts (IBM Deutschland 2022).

Judges provide unique insights into the current state of the art regarding technology use within courts, and they are the target group of the AI systems under consideration. Therefore, their perspectives serve as a critical touchstone for understanding the potential implications arising from AI’s integration into legal proceedings. Building on this premise, through the conduct of 20 in-depth interviews with German judges, this explorative research aims to shed light on the following research questions:

- **RQ1:** According to the surveyed judges, how will a court proceeding appear in the year 2040, and to what extent is AI expected to be involved?
- **RQ2:** What implications would an increased use of AI have on the role of judges and the principle of judicial independence?
- **RQ3:** Based on responses from the judges, should the results and decisions of an AI system prevail over human judgments if the system demonstrably arrives at superior verdicts?

The findings demonstrate a general expectation for the ongoing digitalization of courts, while scenarios for the implementation of AI are only partially conceivable. Concerning the impact of AI on judicial independence, contrasting views were prevalent. Many individuals hold reservations about fully delegating decision-making to machines, perceiving it as both inconceivable and worrisome. Conversely, a portion of respondents advocate for machine-mediated decision-making, contingent upon substantiated evidence demonstrating its superior decision-making capabilities. Overall, the perspectives and views of the surveyed judges are diverse and a clear trend cannot be determined. However, there exists a tendency to evaluate AI implementations in the judiciary more optimistically and positively rather than critically.

**AI in the judiciary: state of the art**

In many countries, it is expected that AI use in legal proceedings will increase in the future. This sentiment is exemplified in China, where an extensive network of AI applications is set to be deployed by 2025, designed to bolster and streamline legal processes (Yu 2022). The United Nations Educational, Scientific and Cultural Organization envisions a rising adoption of AI in the judiciary, evident in the development of a dedicated online course titled “AI and the Rule of Law: Capacity Building for Judicial Systems” (UNESCO 2023). Additionally, a growing demand for the judiciary to go digital has been fueled by citizens’ higher expectations, increased court workloads, succession challenges, and the need to balance the playing field with legal tech providers (IBM Deutschland 2022).

However, gazing into the future is unnecessary, as AI systems are already being formally employed by judges. Risk assessment tools are perhaps best recognized: The objective of these tools is to determine the prospective likelihood of recidivism among offenders. In 49 out of 50 US states, such systems are applied to assess aspects like bail, parole, pretrial custody status, or the duration of sentences (Stevenson 2018). The Chinese AI-driven system ‘Little Judge Bao’ goes further, proposing tailored sentences based on pre-selected factors (Shi 2022). Looking at the state of digitalization, Singapore serves as a notable instance of a highly digitized judiciary with an all-encompassing online case management system across jurisdictions, facilitating case initiation, monitoring, and data for predictive caseload analysis. Canada showcases another example, launching its first online tribunal in 2012, where all court interactions occur digitally (Hartung et al. 2022).

Germany’s judiciary has fallen behind both internationally and compared to other sectors in adopting digital transformation. According to Hartung et al. (2022), the technological solutions implemented within the German judicial system are limited, outdated, and not sufficiently aligned with user requirements. They estimate that the digitalization of the German judiciary lags behind leading countries by approximately 10–15 years. Dreyer and Schmee (2019) conclude that the feasibility of AI in the judiciary fails solely due to the insufficient availability of training data. Despite this lag in digitalization, the use of algorithms in courts is subject of a critical debate among legal scholars. This encompasses discussions on how algorithms could be deployed within the judiciary to address the shortcomings of human decision-making (Nink 2021), the legal evaluation of so-called ‘robot judges’ (Greco 2021), or the implications of AI deployment on human rights (Završnik 2020). In the field of information systems, the topic has received less attention thus far. Some studies examine the recidivism prediction algorithms already in use in the United States, focusing on aspects such as fairness and reliability (Berk 2019), or the effects of human-machine interaction in this context (Grgić-Hlača et al. 2019).
target audience of AI systems in the judiciary is predominantly not directly involved in these studies.

Research method

Sample characteristics
The sample (n = 20) was recruited through email invitations to courts, to the Deutscher Richterbund (German association of judges), and through personal networks. It comprises eleven male and nine female individuals, with an average experience of 13.6 years as judges (sd = 10.3 years). Almost all participants hold active judge positions, with only one individual having ceased working as a judge in 2017. The distribution across judicial levels includes 10 judges from local courts, 8 from regional courts, and 2 from higher regional courts. Regarding age, one of the participants is below 30 years old, seven are between 30 to 39 years old, five are between 40 to 49 years old, five are between 50 to 59 years old, and two are above 60 years old. Nine individuals specialize in civil law, three in criminal law, with four of them holding active judge positions in both civil and criminal law. Two participants each practice administrative law and labor law. Participants responded to the Affinity for Technology Interaction (ATI) Scale (Franke et al. 2019). The results, obtained on a scale ranging from 1 (low affinity) to 6 (high affinity), reveal that, as a group, participants demonstrate a moderate level of affinity for technology interaction, with a mean score of m = 3.47 (sd = .94, range: 2.00–5.00) and a high internal consistency (α = .92). This suggests that the sample is not biased by a strong affinity for technology, which could have been possible since participation in the interviews was voluntary, implying an inherent interest in the topic.

Conducting the interviews
This study adopted the reporting framework, guidelines, and dramaturgical model proposed by Myers and Newman (2007) for conducting interviews within the context of information system research. They suggest incorporating essential meta-data related to the interviews (see table 1).

The initial interview was conducted in person, while all subsequent interviews were conducted virtually. The guiding questionnaire consisted of 29 questions, categorized into six sections: current technology usage, AI system requirements, personal attitudes, expectations, human judges’ capacity, and ethics. The present paper emphasizes the questions within the expectations and ethics categories.

Nearly the entire interview process was audio-recorded. In the initial twelve interviews, recording was omitted for the categories human judges’ capacity and ethics, opting for written notes instead. This approach was intended to foster greater trust and enhance participants’ confidence, given the sensitive nature of these questions. However, this method did not produce the desired outcome. As a result, for the subsequent eight interviews, the entire interview process was recorded. The content analysis that followed used the MAXQDA software.

Qualitative analysis
The content analysis was guided by the methodological frameworks put forth by Kuckartz and Rädiker, encompassing both their general approaches and the specific techniques employed for analyzing interviews (Kuckartz and Rädiker 2019; Rädiker and Kuckartz 2020). For the development of the coding scheme, a data-driven approach (inductive methodology) was adopted, involving a step-by-step coding process where codes were iteratively generated until saturation was achieved. The aim is to structure the content and analyze it on the basis of this structure. In this way, diverse attitudes and opinions can be identified. The two authors initially independently coded three randomly selected interviews. The results were then discussed and harmonized. Subsequently, the remaining interviews were independently coded, and their outcomes, such as their alignment with existing codes or the creation of new codes, were deliberated upon. The overall categorical system consists of 72 codes and a total of 339 text passages were coded.

Results

The judiciary in 2040
The judges were asked how a court proceeding might look in the year 2040. The question was received differently, yielding a wide range of responses, as indicated by the numerous codes generated (25 in total). These responses can be categorized into two main themes: future scenarios that describe expectations and desired outcomes. As a result, for the subsequent eight interviews, the entire interview process was recorded. The content analysis that followed used the MAXQDA software.

Qualitative analysis
The content analysis was guided by the methodological frameworks put forth by Kuckartz and Rädiker, encompassing both their general approaches and the specific techniques employed for analyzing interviews (Kuckartz and Rädiker 2019; Rädiker and Kuckartz 2020). For the development of the coding scheme, a data-driven approach (inductive methodology) was adopted, involving a step-by-step coding process where codes were iteratively generated until saturation was achieved. The aim is to structure the content and analyze it on the basis of this structure. In this way, diverse attitudes and opinions can be identified. The two authors initially independently coded three randomly selected interviews. The results were then discussed and harmonized. Subsequently, the remaining interviews were independently coded, and their outcomes, such as their alignment with existing codes or the creation of new codes, were deliberated upon. The overall categorical system consists of 72 codes and a total of 339 text passages were coded.

Results

The judiciary in 2040
The judges were asked how a court proceeding might look in the year 2040. The question was received differently, yielding a wide range of responses, as indicated by the numerous codes generated (25 in total). These responses can be categorized into two main themes: future scenarios that describe expectations and desired outcomes. As a result, for the subsequent eight interviews, the entire interview process was recorded. The content analysis that followed used the MAXQDA software.

Qualitative analysis
The content analysis was guided by the methodological frameworks put forth by Kuckartz and Rädiker, encompassing both their general approaches and the specific techniques employed for analyzing interviews (Kuckartz and Rädiker 2019; Rädiker and Kuckartz 2020). For the development of the coding scheme, a data-driven approach (inductive methodology) was adopted, involving a step-by-step coding process where codes were iteratively generated until saturation was achieved. The aim is to structure the content and analyze it on the basis of this structure. In this way, diverse attitudes and opinions can be identified. The two authors initially independently coded three randomly selected interviews. The results were then discussed and harmonized. Subsequently, the remaining interviews were independently coded, and their outcomes, such as their alignment with existing codes or the creation of new codes, were deliberated upon. The overall categorical system consists of 72 codes and a total of 339 text passages were coded.

Results

The judiciary in 2040
The judges were asked how a court proceeding might look in the year 2040. The question was received differently, yielding a wide range of responses, as indicated by the numerous codes generated (25 in total). These responses can be categorized into two main themes: future scenarios that describe expectations and desired outcomes. As a result, for the subsequent eight interviews, the entire interview process was recorded. The content analysis that followed used the MAXQDA software.

Qualitative analysis
The content analysis was guided by the methodological frameworks put forth by Kuckartz and Rädiker, encompassing both their general approaches and the specific techniques employed for analyzing interviews (Kuckartz and Rädiker 2019; Rädiker and Kuckartz 2020). For the development of the coding scheme, a data-driven approach (inductive methodology) was adopted, involving a step-by-step coding process where codes were iteratively generated until saturation was achieved. The aim is to structure the content and analyze it on the basis of this structure. In this way, diverse attitudes and opinions can be identified. The two authors initially independently coded three randomly selected interviews. The results were then discussed and harmonized. Subsequently, the remaining interviews were independently coded, and their outcomes, such as their alignment with existing codes or the creation of new codes, were deliberated upon. The overall categorical system consists of 72 codes and a total of 339 text passages were coded.

Results

The judiciary in 2040
The judges were asked how a court proceeding might look in the year 2040. The question was received differently, yielding a wide range of responses, as indicated by the numerous codes generated (25 in total). These responses can be categorized into two main themes: future scenarios that describe expectations and desired outcomes. As a result, for the subsequent eight interviews, the entire interview process was recorded. The content analysis that followed used the MAXQDA software.

Qualitative analysis
The content analysis was guided by the methodological frameworks put forth by Kuckartz and Rädiker, encompassing both their general approaches and the specific techniques employed for analyzing interviews (Kuckartz and Rädiker 2019; Rädiker and Kuckartz 2020). For the development of the coding scheme, a data-driven approach (inductive methodology) was adopted, involving a step-by-step coding process where codes were iteratively generated until saturation was achieved. The aim is to structure the content and analyze it on the basis of this structure. In this way, diverse attitudes and opinions can be identified. The two authors initially independently coded three randomly selected interviews. The results were then discussed and harmonized. Subsequently, the remaining interviews were independently coded, and their outcomes, such as their alignment with existing codes or the creation of new codes, were deliberated upon. The overall categorical system consists of 72 codes and a total of 339 text passages were coded.

Results

The judiciary in 2040
The judges were asked how a court proceeding might look in the year 2040. The question was received differently, yielding a wide range of responses, as indicated by the numerous codes generated (25 in total). These responses can be categorized into two main themes: future scenarios that describe expectations and desired outcomes. As a result, for the subsequent eight interviews, the entire interview process was recorded. The content analysis that followed used the MAXQDA software.

Qualitative analysis
The content analysis was guided by the methodological frameworks put forth by Kuckartz and Rädiker, encompassing both their general approaches and the specific techniques employed for analyzing interviews (Kuckartz and Rädiker 2019; Rädiker and Kuckartz 2020). For the development of the coding scheme, a data-driven approach (inductive methodology) was adopted, involving a step-by-step coding process where codes were iteratively generated until saturation was achieved. The aim is to structure the content and analyze it on the basis of this structure. In this way, diverse attitudes and opinions can be identified. The two authors initially independently coded three randomly selected interviews. The results were then discussed and harmonized. Subsequently, the remaining interviews were independently coded, and their outcomes, such as their alignment with existing codes or the creation of new codes, were deliberated upon. The overall categorical system consists of 72 codes and a total of 339 text passages were coded.
to “better process documents with computers” are anticipated. Moreover, parties are expected to “communicate electronically with each other”, “submit a relatively significant amount electronically”, and “hopefully, we will indeed have an electronic case file”. Digital documentation is expected to increase, for example, due to the recording of hearings, and it is anticipated that evidence will be technologically processed and digitally accessible.

- **Video hearing (14):** The judges are confident that video conferencing for hearings will continue to expand, anticipating increased opportunities in this regard.

- **Hardly any change (7):** Some judges do not anticipate significant changes within the judiciary, and if any changes are foreseen, they are mostly limited to minor developments in digitalization.

- **AI deployment (5):** Certain judges anticipate an increased use of AI, for instance, in tasks such as suggesting applicable norms for a case, organizing precedent cases, automating the creation of basic legal documents, and employing predictive analysis tools.

- **Pre-trial proceedings (5):** Some judges anticipate the establishment of pre-trial proceedings such as the implementation of predictive tools, resulting in the avoidance of less promising lawsuits, the use of algorithms developed by companies or other private sector stakeholders to decide specific disputes, and the existence of online dispute resolutions.

In addition, the following aspects were addressed, which characterize the scenarios in more detail:

- **Decision-making remains with humans (4):** Four respondents are certain that the ultimate decision will remain with the judge, as one individual expressed, “I believe we still need judges who make the final decision”.

- **Reduced processing times (4):** It is anticipated that, among other factors, the implementation of new technologies will enable court proceedings to be conducted in a shorter span of time.

- **Change in organizational structure (3):** An “infrastructure reform” is anticipated, and it is also assumed that there will be fewer courts and judges as well as fewer case numbers, particularly in civil law.

Furthermore, the following statements were mentioned once each: the potential for new citizen-court interactions, such as online lawsuit filings; the potential partial replacement of judges; the potential use of Virtual Reality; the emergence of digital lawyers for defendants; and the anticipated collaboration enhancement within the EU, possibly facilitated through an EU-wide shared database for decisions.

Regarding the mentioned concerns and critical considerations, it was noted six times that face-to-face conversation is irreplaceable, and four times it was emphasized that human interaction cannot be substituted, with one person saying: “What distinguishes judicial decisions and court proceedings at their core, however, is the personal conversation and the individual context within a legal process. I believe that this cannot be replaced by AI systems because there is a significant amount of social interaction involved, which may not directly relate to legal matters but nonetheless significantly shapes the situation.” Two individuals stated that they believe older judges will struggle with the growing digitalization. Another two highlighted the importance of a societal debate about the use of AI in the legal system, questioning whether we as a society desire such developments. Two respondents expressed concerns about the increasing reliance on technology. The following concerns were raised once each: the growing digital asymmetry within the legal profession, IT security, the lack of competence of IT service providers, and concerns about the rule of law.

### The future role of judges

Subsequently, participants were asked about their expectations regarding the development of the judge’s role by 2040. The responses varied between positive expectations, concerns, and neutral statements (see table 2).

<table>
<thead>
<tr>
<th>Optimistic Anticipation</th>
<th>£</th>
<th>Concerns</th>
<th>£</th>
<th>Neutral</th>
<th>£</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relief through digitalization</td>
<td>7</td>
<td>Reduced decision-making authority</td>
<td>2</td>
<td>No changes of the judge’s role</td>
<td>6</td>
</tr>
<tr>
<td>AI as support and assistance</td>
<td>4</td>
<td>Reduced reverence</td>
<td>1</td>
<td>Judges as case managers and mediators</td>
<td>5</td>
</tr>
<tr>
<td>AI in mass proceedings</td>
<td>1</td>
<td>Rise in information overflow</td>
<td>1</td>
<td>Judgment remains with the human</td>
<td>3</td>
</tr>
<tr>
<td>Additional responsibilities</td>
<td>1</td>
<td>New competencies necessary</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Surveillance of the systems</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Tab. 2:** Categorization of potential changes of the judge’s role by 2040.

*Source: interviews conducted by the authors*
The question was raised as to whether the implementation of AI systems within the judiciary might lead judges to excessively rely on them, potentially fostering automation bias – a tendency to overly trust automated systems, which may result in errors or overlooking something (Skitka et al. 2000). This question specifically pertained to decision support systems, where the human makes the final decision, for instance, pre-filing court orders. According to Sheridan’s automation scale – ranging from 1 = the human must decide and execute everything to 10 = the system acts autonomously and decides without human involvement – these systems fall within levels up to a maximum of 5 (Sheridan et al. 1978).

Some judges noted this issue persists even without AI, such as when they agree to a prosecutor’s case dismissal request to reduce workload (mentioned in 4 interviews). Younger individuals, with greater trust in technology, were identified by two respondents as more prone to agree with the system, reinforcing automation bias. Additionally, it was noted that judges often face time constraints, which could lead them to go along with the system’s decision simply due to time pressure (4). Proposed solutions included the need for judges to receive appropriate training (1), designing the systems and their usage context with psychological incentives to avoid automation bias (1), and the implementation of relevant regulations (1). In contrast, some respondents (5) believe that automation bias is not a problem for judges because they are “self-disciplined”, that they have “inherent skepticism towards anything that touches their own high decision-making authority”, and, ultimately, that the “professional group is inherently inclined to resist”.

Participants were also asked whether judicial independence is called into question with an increased use of AI. On the one hand, some (5) believe the development to be critical due to concerns about a gradual takeover by such systems, with judges increasingly facing pressure to justify decisions that do not align with those of AI, and fears that AI might render judges redundant. On the other hand, the argument was made that judicial independence is not at risk, as ultimately, judges decide how and when to use technology, with AI systems serving merely as assistants (7). Moreover, it was emphasized that a threat to judicial independence would depend on whether the use of AI systems would be mandatory and how the integration of such algorithms into procedural rules would occur (12). Additionally, it was stressed twice that it is within the responsibility of individual judges to determine whether their own independence would be compromised or not: “I believe that an AI system can pose a significant threat to lazy-minded judges.”

Perspectives on AI-generated judgments
During the interviews, the judges were also asked: Should AI system outcomes override human judgments when the system consistently yields better verdicts? In this context, the discussion pertained to systems classified at level 10 on the automation scale, meaning they operate autonomously without human intervention (Sheridan et al. 1978). The majority of respondents (14) initially countered by stating that it is not demonstrable at what point a decision would be considered “better”. Subsequently, responses to this question diverged significantly. Some supported the idea of machines issuing judgments, while others endorsed it only under certain conditions or for specific use cases. Conversely, many responses entailed explicit and absolute rejection (see table 3). As evident from the frequency of mentions, it is apparent that not only were the respondents divided in their opinions, but individual participants also provided varying statements. Three statements not included in the table indicate that the responsibility of AI systems for decision-making is a societal choice: “I believe that, since we live in a democracy, if society decides that we want this, it should be done.”

Discussion
Judges operate with autonomy, determining their operational methodologies, and hold accountability for each procedural facet, as articulated in Art. 97 Abs. 1 GG, which underscores their independence and subordination solely to the law. They typically lack dedicated secretarial support or personally assigned assistants, exemplifying the self-directed nature of their role. Consequently, this engenders, on the one hand, the fundamental latitude for judges to exercise discretion in adopting technologi-

<table>
<thead>
<tr>
<th>Perspective</th>
<th>Argument</th>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approval</td>
<td>Human beings prone to errors</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>If demonstrably superior judgments, then approval</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>The machine is more powerful</td>
<td>1</td>
</tr>
<tr>
<td>Conditional approval</td>
<td>Usage allowed but adopting the results not mandatory</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Usage if case-by-case justice appropriate</td>
<td>2</td>
</tr>
<tr>
<td>Approval for specific use cases</td>
<td>AI usage for mass proceedings</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>AI in preceding administrative actions</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>In some cases conceivable (without specifying)</td>
<td>1</td>
</tr>
<tr>
<td>Rejection</td>
<td>Human perception and responsibility crucial</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>End of judicial independence</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Hierarchy of instances rendered obsolete</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Rule of law concerns</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Training data susceptible to manipulation</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>AI verdict not accepted by humans</td>
<td>1</td>
</tr>
</tbody>
</table>

Tab. 3: Categorization of responses to the question of whether AI machines should render judgments instead of humans. Source: interviews conducted by the authors
tal aids, unless statutory provisions dictate otherwise. On the other hand, it underscores the challenge of seamlessly integrating AI-based assistance systems into existing judicial processes. It is to be noted that the interviews do not reveal a consistent consensus; a multitude of diverse viewpoints were expressed. However, it is noticeable that there is a general tendency towards a more favorable outlook rather than a critical one regarding the future possibilities of AI application, such as its potential use as a helpful tool in mass proceedings or as an assistance system for case processing. At the same time, critical topics such as IT security or data for such systems and the associated potential for discrimination were scarcely addressed. The slightly positive view held by judges could be attributed to their potential lack of AI expertise compared to members of, for instance, the information systems community, leading to a limited understanding of the technological challenges associated with AI. Judges tend to perceive digitalization and related AI technologies as advantageous for their daily work, hence the positive outlook.

Regarding the first research question concerning expectations for the year 2040, the responses demonstrate a strong reliance on digitalization. This underscores the previously mentioned lag in the German judiciary and the judges’ expectations that this gap will be bridged in the coming years. Increased AI deployment is only expected to a limited extent. Regarding the second research question concerning potential impacts on the role of judges, on the one hand, a positive expectation was revealed, such as relief through digitalization. On the other hand, concerns were expressed, for instance, regarding reduced decision-making authority.

A similar pattern emerged in response to the question about judicial independence, with some expressing concerns about a gradual takeover by AI, while others had no reservations. The risk of automation bias coming with regular use of such systems, in turn, was largely acknowledged. Concerning the third research question about whether judgments from AI systems should potentially be deemed more significant than those made by humans, there were supporters who could envision such a scenario under specific circumstances, as well as opponents who assert that such an outcome is precluded. Notably, the diversity of responses to the previous questions remained evident in addressing this question as well, although there is research that demonstrate that higher levels of automation are frequently met with less acceptance when compared to lower levels (Ghazizadeh et al. 2012).

The study has notable limitations to consider. It is confined to the German context, potentially impacting its applicability to other legal systems. Due to the small sample size, the study is not representative. Also, the judges’ self-selection might introduce bias, as they hold an interest in AI. Finally, different interpretations and confusion regarding AI and digitalization were observed. Despite the predominantly descriptive nature of the analysis, it might serve as a valuable resource for future research endeavors, particularly for theory building.

**Outlook**

According to the draft of the EU AI Act (Article 8 of Annex III), AI systems used in the judiciary are classified as high risk. Therefore, the deployment of AI systems for judges is already politically viewed with skepticism. As a result, scenarios involving the use of AI give rise to various legal, technical, and ethi-

*Nearly one-third of the respondents anticipate that the role of the judge will not change.*

...


ANNA-KATHARINA DHUNGEL holds a Master’s degree in International Politics and International Law, as well as a Master in Business Informatics. With five years of experience as an IT Consultant, she joined the University of Lübeck in 2021 to pursue her PhD and conduct research on the implementation of artificial intelligence in the field of justice.

PROF. DR. MOREEN HEINE has been a professor of E-Government and Open Data Ecosystems at the University of Lübeck and scientific director of the Joint Innovation Lab since 2019. Previously, she was an assistant professor of business information systems at the University of Potsdam. She is a National E-Government Competence Centre (NEGZ) board member.
AI and access to justice: How AI legal advisors can reduce economic and shame-based barriers to justice

Brandon Long*, 1, Amitabha Palmer 2

Abstract • ChatGPT – a large language model – recently passed the U.S. bar exam. The startling rise and power of generative artificial intelligence (AI) systems such as ChatGPT lead us to consider whether and how more specialized systems could be used to overcome existing barriers to the legal system. Such systems could be employed in either of the two major stages of the pursuit of justice: preliminary information gathering and formal engagement with the state’s legal institutions and professionals. We focus on the former and argue that developing and deploying publicly funded AI legal advisors can reduce economic and shame-based cultural barriers to the information-gathering stage of pursuing justice.

Keywords • artificial intelligence, shame, barriers to justice, philosophy of technology, law

This article is part of the Special topic "AI for decision support: What are possible futures, social impacts, regulatory options, ethical conundrums and agency constellations?," edited by D. Schneider and K. Weber. https://doi.org/10.14512/tatup.33.1.21

Introduction

In most countries, the legal system is the primary institution through which citizens pursue justice when wronged or to exercise their rights. However, in many of these countries, a variety of barriers prevent citizens from accessing the legal system (OECD 2015). The primary barriers are economic, cultural, and political and they generate significant externalities such as social exclusion, government dependence, weaker business assurances, and higher healthcare costs (OECD 2015, p. 4). It is not only a just but a prudent state that seeks to reduce or eliminate barriers to justice.

ChatGPT – a large language model (LLM) – recently passed the U.S. bar exam (Arredondo et al. 2023). The startling rise and power of generative AI systems such as ChatGPT lead us to consider whether and how more specialized systems could be employed to overcome existing barriers to the legal system. Broadly, such systems could be employed in either of two major stages of the pursuit of justice: preliminary information gathering and formal engagement with the State’s legal institutions and professionals. We focus on the former and argue developing and deploying publicly funded artificial intelligence legal advisors...
(AI LAs) can reduce economic and shame-based cultural barriers to the information-gathering stage of pursuing justice.

By AI LA we have in mind a system that provides potential litigants with reliable legal information that is specific and intelligible to allow them to make informed decisions whether to formally contract a lawyer and/or formally pursue their claims in court. Several similar legal AI platforms already exist (AI Lawyer 2023; Rattray 2023; Casetext 2022) and will likely only improve over time. In one London-based law firm from November 2022 to February 2023, Harvey AI was used by 3,500 of their lawyers to ask 40,000 legal questions during their day-to-day work (Rattray 2023). Specialized AI models can already give advice for specific domains of law. For example, JusticeBot (Tribunal administratif du logement 2023), a free tool for Quebec housing law, takes case facts into account in giving legal advice, asks pertinent questions, and cites similar cases for each relevant legal claim.

Throughout, we do not claim that highly reliable AI LAs currently exist but assume they will be available soon. As such, a foresighted technology assessment should consider their use before such technologies are developed and become available. We advocate that AI LAs be publicly funded. While a privately developed AI LA could achieve the same technical goals, a publicly funded AI LA supports broader economic accessibility. Moreover, democratic governments and international organizations should support implementation of AI LAs. Access to justice is intrinsically good but also provides instrumental benefits. It is a crucial prerequisite for establishing legal confidence and trust, thereby creating a favorable business environment, attracting investments, and contributing to overall economic spending (The Perryman Group 2009, pp. 19–21). Growing evidence suggests that the ability to address legal issues and obtain justice has a positive impact on inclusive economic growth (OECD 2013, p. 2, 2015, pp. 1–4). This impact is manifested through job creation, reduced absences at work due to legal problems (Task Force on Justice 2019, p. 45), improved housing stability, resolution of debt, and stimulating growth by instilling confidence in assurance (Stolper et al. 2007, pp. 8–9). Equal access to justice may also foster economic growth by establishing a level playing field (Task Force on Justice 2019, pp. 39–41), especially for small or medium economic participants (OECD 2015, pp. 3–4). It also facilitates enforcement of contracts, encourages fair competition, and instills confidence in regulatory frameworks (OECD 2015, pp. 9–10). Thus, supporting access to justice can play a role in assisting individuals to overcome severe forms of social exclusion and ensuring equal opportunities for economic advancement.

To support our thesis that AI LAs can reduce barriers to justice, we (1) outline common economic and shame-based cultural barriers to pursuing legal justice. (2) describe how an AI LA can mitigate barriers during the information-gathering stage, and (3) address potential limitations and harms. Our scope for these claims is Anglo-American common law systems. This brings with it unique barriers to legal aid and implementing AI systems, and a specific common law that may or may not generalize more globally.

### Economic barriers

#### Economic barriers to legal aid seeking

This section reviews economic barriers to justice and suggests how an AI LA could reduce them during the information-gathering stage. Economic barriers are not only financial, but also the opportunity cost of time spent on information-seeking and transportation.

A substantial body of evidence finds people with low socio-economic status (SES) face greater barriers to the legal system – and therefore they also face greater barriers to gain access to justice (Commission on Legal Empowerment of the Poor 2008, pp. 6–9; Legal Services Corporation 2022, sec. 5; OECD 2015, p. 7). Poverty, poverty-related discrimination, and distrust present barriers to justice globally (Beqiraj and McNamara 2014, chaps. 4–5). What is more, marginalized – including economically marginalized – populations face unique barriers to justice in the UK (Gill et al. 2021) and in Canada (Silverman and Molnar 2016).

Financial barriers influence whether people pursue justice through the legal system. Across OECD countries, 42% to 90% of individuals who opt out of pursuing legal aid attribute their decision to financial considerations, whether real or perceived (OECD 2015, p. 5). Further, for 92% of legal problems low-income Americans face, they do not receive any or enough legal aid (Legal Services Corporation 2022, pp. 47–48). Moreover, education and accessible information are also barriers to the justice system – 53% of low-income Americans do not know if they are able to find an affordable lawyer if needed (Legal Services Corporation 2022, pp. 51–52). These and other barriers lead them to pursue litigation at lower rates than higher SES groups. For example, in medical contexts, lower SES groups pursue litigation at lower rates when compared to other groups because of a lack of access to legal resources and the nature of the contingency fee system in medical malpractice claims (McClellan et al. 2012; Viser 2022).

Our thesis is that artificial intelligence legal advisors can reduce barriers to justice.
Overcoming economic barriers with publicly funded AI legal advisors

To specify ways in which AI technology can reduce economic barriers to justice we propose conceiving the pursuit of legal justice as having two stages:

1. Information gathering: In this stage, one seeks to determine whether one has a claim, evaluate the strength of that claim, and evaluate the cost-benefit tradeoffs of formally pursuing one’s legal claim.

2. Formal engagement: In this stage, one hires a lawyer and engages with state officials and the court system to pursue one’s claim.

People with limited economic resources cannot frivolously engage with an economically onerous legal system. Before one invests resources to make an informed formal pursuit of a claim, one must have some sense of one’s prospects for success and the underlying legal reasoning. Hence, existing economic barriers to information gathering prevent people who, unbeknownst to them, have strong claims and might have pursued them formally had they possessed this knowledge. We believe AI LAs are well-suited to address economic barriers to information gathering.

We have in mind an AI LA that could provide prospective litigants with (a) an assessment of legal considerations and reasoning involved in their claim, (b) a crude assessment of their case’s likelihood of success in court, (c) an interactive lay explanation of (a) and (b).

Assessment of legal considerations would include explanations of which laws apply, why and how they apply, and how similar cases have been treated. The crude assessment of the likelihood of legal success would be expressed as ‘poor,’ ‘unlikely,’ ‘unknown,’ ‘fair,’ or ‘strong.’ Critically, like existing LLMs, AI advisors will be conversational, allowing users to ask follow-up questions and clarifications.

Citizens with a limited understanding of their legal situation, the likelihood of success, and scarce economic resources may be hesitant to approach a lawyer. Proposed capabilities for the AI LA align with the information citizens seek during information gathering. Unlike consulting a lawyer, a government-funded AI LA can provide legal information sought without imposing burdensome financial, time, and transportation costs. Moreover, if it is publicly funded, citizens bear minimal direct costs and online accessibility eliminates transportation and mitigates time costs.

Under this model, citizens who otherwise might not have pursued legitimate claims due to economic barriers in the information-gathering stage may now choose to pursue them. Nevertheless, economic barriers are not the only barriers to justice. We now turn to investigating how legal AI can address cultural and shame-based barriers to justice.

Cultural barriers to justice

In this section, we (a) define shame and how it relates to cultural norms, (b) explain how it can pose a barrier to pursuing justice in the information-gathering stage, (c) suggest how a publicly funded AI LA can mitigate barriers to legal information seeking. Notice in the following that while economic barriers to justice may be addressed by funding human legal resources, AI LAs have unique features that address shame-based barriers in ways additional funding cannot.

Shame, stigma, culture

Shame is a “negative emotion that arises when one is seen and judged by others (whether they are present, possible or imagined) to be flawed in some crucial way, or when some part of oneself is perceived to be inadequate, inappropriate or immoral” (Dolezal and Lyons 2017, p. 257). Shame influences behavior because it can threaten one’s feelings of belonging and acceptance within interpersonal contexts, socially, and politically (Walker and Bantebya-Kyomuhendo 2014).

Shame can be acute or chronic. Acute shame is a single episode that arises unexpectedly, as in cases of embarrassment where in social interaction, one’s self-presentation falters, fails, or falls short of socially desired modes of comportment (Dolezal and Lyons 2017). Chronic shame is often a result of general social stigma directed at marginalized social groups. For instance, shame is linked to racism, discrimination (Harris-Perry 2011),

A government-funded artificial intelligence legal advisor can provide legal information sought without imposing burdensome costs.

‘unknown,’ ‘fair,’ or ‘strong.’ Critically, like existing LLMs, AI advisors will be conversational, allowing users to ask follow-up questions and clarifications.

Citizens with a limited understanding of their legal situation, the likelihood of success, and scarce economic resources may be hesitant to approach a lawyer. Proposed capabilities for the AI LA align with the information citizens seek during information gathering. Unlike consulting a lawyer, a government-funded AI LA can provide legal information sought without imposing burdensome financial, time, and transportation costs. Moreover, if it is publicly funded, citizens bear minimal direct costs and online accessibility eliminates transportation and mitigates time costs.

Under this model, citizens who otherwise might not have pursued legitimate claims due to economic barriers in the information-gathering stage may now choose to pursue them. Nevertheless, economic barriers are not the only barriers to justice. We now turn to investigating how legal AI can address cultural and shame-based barriers to justice.

Cultural barriers to justice

In this section, we (a) define shame and how it relates to cultural norms, (b) explain how it can pose a barrier to pursuing justice in the information-gathering stage, (c) suggest how a publicly funded AI LA can mitigate barriers to legal information seeking. Notice in the following that while economic barriers to justice may be addressed by funding human legal resources, AI LAs have unique features that address shame-based barriers in ways additional funding cannot.

Shame, stigma, culture

Shame is a “negative emotion that arises when one is seen and judged by others (whether they are present, possible or imagined) to be flawed in some crucial way, or when some part of oneself is perceived to be inadequate, inappropriate or immoral” (Dolezal and Lyons 2017, p. 257). Shame influences behavior because it can threaten one’s feelings of belonging and acceptance within interpersonal contexts, socially, and politically (Walker and Bantebya-Kyomuhendo 2014).

Shame can be acute or chronic. Acute shame is a single episode that arises unexpectedly, as in cases of embarrassment where in social interaction, one’s self-presentation falters, fails, or falls short of socially desired modes of comportment (Dolezal and Lyons 2017). Chronic shame is often a result of general social stigma directed at marginalized social groups. For instance, shame is linked to racism, discrimination (Harris-Perry 2011),
tizing theft or domestic violence isn’t a bad cultural practice. However, we suggest shame and shame-inducing norms that impede the legitimate pursuit of justice are prima facie bad and should be eliminated. These include norms against litigation due to group membership and norms against disclosing information that stigmatizes or shames.

In the cases below, we show how AI LAS can mitigate shame-based barriers to justice due to the human propensity not to feel judged when interacting with AI (Bartneck et al. 2010; Holthöwer and Van Doorn 2023).

Overcoming shame-based cultural barriers with publicly funded AI legal advisors

First case: Shame-based barriers to justice for victims of intimate partner violence

Victims of intimate partner violence (IPV) often forgo pursuing justice because the stigmatization of being a victim can lead to shame (Overstreet and Quinn 2013). IPV survivors grapple with a lasting sense of shame after their experience, stemming from lost self-identity, self-blame, and fear of judgment (Camp 2022, p. 103). Seeking help often leads to encounters with people or institutions— including the legal system—that worsen rather than alleviate this shame (Camp 2022, pp. 136–137). Understandably, individuals who perceive a stigma associated with being a victim of IPV are less likely to seek institutional support, and when they do disclose experiences, they prefer indirect language that hints at abuse without disclosing details (Williams and Mickelson 2008).

The most common legal intervention for victims of IPV is protection orders, vital tools for responding to IPV. Yet, obtaining a protection order requires survivors to enter “a process that often deprives them of their privacy and ability to control their self-image—experiences anchored in shame” (Camp 2022, pp. 103–104). This suggests shame may both be a barrier to disclosing information and for seeking aid. For example, one U.S. abuse shelter from 2004–2008 found of all victims of IPV, only 32.25% had protection orders upon appearing at the shelter (Durfee and Messing 2012).

AI LAS can mitigate shame-based barriers to information gathering for victims of IPV. Interacting with an AI rather than a human restores privacy and eliminates shame that can be induced by the presence of others which can allow a victim of IPV to safely learn (a) what legal recourse and protections are available to them, (b) how to pursue legal recourse/protection, (c) whether their circumstances meet legal criteria, (d) the likelihood they will succeed in obtaining legal recourse or protection, and (e) all the above in an interactive lay-friendly language. Certainly, victims of IPV will have to engage with humans if they decide to pursue recourse formally. However, AI LAS allow the acquisition of information for an informed legal decision. Moreover, since social stigmas associated with being a victim of IPV are unjustified and harmful, access to an AI LA justifiably reduces shame-based harms to victims of IPV.

Second case: Shame-based barriers to justice for marginalized groups when cultural norms obscure legal rights

Cultural norms may stigmatize seeking legal aid for women or people in positions of lower social status where fear of reprisal or shame keeps legal complaints underserved (Long Chamness and Ponce 2019, pp. 13–17). This may be common outside the U.S., where compensation culture is weak, nonexistent, or displaced by other norms in specific contexts. Consider the following case involving inheritance rights.

In some communities, there is a cultural expectation that women will relinquish their inheritance rights to their brothers when their parents die (Nayeen 2020). In doing so, a woman protects and ensures their culturally coveted status of a ‘good sister’. Conversely, failure to relinquish her right puts her social status in jeopardy and incurs stigma from norm violation.

The prevalence of a cultural norm for women to relinquish their inheritance rights can create confusion about what legal inheritance rights women have. Moreover, it can prevent women from inquiring about their rights as this could be interpreted as a pre-emptive norm violation (Nayeen 2020). Even inquiring into one’s rights can incur stigma or shame—especially if one isn’t sure of the extent of one’s rights or whether one would indeed pursue them.

This cost often prohibits investigating legal rights, which prevents obtaining information required to make informed legal decisions. With full information, a woman may reason the benefit of asserting her inheritance rights offsets the social cost of norm violation. Furthermore, social norms can never be overturned until women living in such communities accurately understand their rights. In short, the conflation between social norms and legal rights deprives women (and other similarly situated marginalized groups) of the opportunity to make informed decisions regarding whether they wish to exercise their inheritance rights.

An AI LA can provide women with the information necessary to make informed decisions regarding tradeoffs between social sanction and exercising inheritance rights (or other rights) is worth it. Such information includes: (a) clarifying any confusion with respect to the nature and extent of the rights in question, (b) other legal variables, (c) the legal process required for exercising rights, (d) a coarse-grained assessment of the case’s likelihood of success in court, and (e) an interactive lay explanation of preceding information.

This is how our model can mitigate cultural barriers to justice in the information-gathering stage. An AI LA permits private inquiries into rights in a way immune to shame. This is most true in small or tight-knit communities where being seen walking into a law office could cause gossip and shame.

Third case: Shame as a barrier to justice for victims of fraud

Disclosing to others that one has been a victim of fraud brings about acute shame that can prevent victims from pursuing justice. In the U.S., for example, fraud is an enormous problem, as consumers lost nearly $9 billion in 2022 (Fair 2023). However, victims rarely come forward and pursue justice. A sur-
A common concern with many AI systems is that they can inherit and reproduce biases in their training data. This concern also applies to AI LAs who will have been trained in case law rife with historical biases. This topic of biases in AI is large and ongoing. An exhaustive treatment goes beyond the scope of this research article. However, a brief response is warranted.

First, the question of biases will always be comparative. There is unlikely ever to exist any human-developed system free of all biases. The question, therefore, is whether an AI LA could have fewer biases than the current system. We believe the answer is ‘yes’ because it is much easier to alter the biases of an AI than it is of the individuals and institutions that compose the entire legal system.

Bias

Bias inhabits AI systems within their training data, algorithms, and outputs. We know that biased data leads to biased algorithms. Therefore, it’s possible to mitigate bias through debiasing the training set or through careful selection of training data. Where this isn’t possible, it’s possible to adjust algorithms that we know were developed using biased data. Finally, if we know in advance that an AI’s outputs are biased, it’s possible to have the AI correct in the other direction (Fazelpour and Danks 2021). While these correction measures are not easy or foolproof, they are easier and more likely to succeed than attempting to correct the implicit and systemic biases of every individual and institution that compose current justice systems. Finally, addressing biases in a legal system can happen concurrently while addressing biased AI LAs in the ways we have mentioned.
Responsibility
AI generates questions about legal responsibility within existing legal frameworks (Beckers and Teubner 2021). An AI LA could make two major kinds of errors that lead to harm which raise questions about responsibility and liability: The AI (a) recommends pursuing litigation when there is no viable claim, or (b) recommends abstaining from litigation when in fact there is a viable claim. The topic of responsibility in AI ethics is rich and complicated and cannot be addressed comprehensively within the constraints of this research article. Nevertheless, a few brief remarks are in order.

In the first case, the issues of responsibility and liability are relatively unproblematic. The AI LA recommends pursuing a claim which leads the user to contact a lawyer. If the AI has erred, the lawyer should explain why further legal action would be inappropriate. If the lawyer is correct, there is no harm save a consultation fee. If the lawyer is incorrect, the lawyer bears the responsibility just as they are currently held responsible for poor legal advice.

In the second case, a fund liability model is appropriate (Beckers and Teubner 2021, pp. 139–140). In this model, a regulatory agency creates and administers a fund or insurance to provide compensation for harm. Firms in the industry sector finance the fund according to their market share and the agency determines ex-post liability for each case/class of cases. Finally, such a model will require that AI LA be audited at appropriate intervals since naïve individuals will not know when the AI’s advice not to litigate is mistaken.

Conclusion
We have explained how economic cost and shame present barriers to accessing the justice system, how AI LAs may alleviate these barriers, and have covered some limitations and harms of such LLMs. There is no one solution to every legal barrier for everyone, but AI LAs present several viable solutions. Such advisors can reduce economic and shame-based barriers to the information-gathering stage of pursuing justice. This is significant since lack of information is itself a barrier to informed decision-making regarding whether to formally pursue justice. We take the value of justice to be intrinsic and self-evident, therefore, expanding access to justice is a good thing. The legal system becomes more just when the cases reaching the court do so based on merit rather than arbitrary barriers.

Funding • This work received no external funding.
Competing interests • The authors declare no competing interests.

References
AI Lawyer (2023): AI lawyer blog. Available online at https://ailawyer.pro/blog, last accessed on 03.01.2024


Legal Services Corporation (2022): The justice gap. The unmet civil legal needs of low-income Americans. Washington, DC: Legal Services Corporation. Available online at https://tsc-live.app.box.com/s/xlv2unuotbzbhwutlj0@emp3my2z, last accessed on 03. 01. 2024.


BRANDON LONG
is currently an M. A. student at Bowling Green State University and will be seeking a PhD after its completion. He is currently working on bioethical arguments for and against genetic enhancement.

AMITABHA PALMER, PHD
holds a PhD in Philosophy and is a HEC-C certified Clinical Ethicist. He is an Instructor and Clinical Ethicist at the University of Texas MD Anderson Cancer Center. His primary areas of research include the ethics of AI, medical ethics, political philosophy, and the effects of medical misinformation on clinical interactions.
Artificial intelligence and judicial decision-making: Evaluating the role of AI in debiasing

Giovana Lopes*

Abstract • As arbiters of law and fact, judges are supposed to decide cases impartially, basing their decisions on authoritative legal sources and not being influenced by irrelevant factors. Empirical evidence, however, shows that judges are often influenced by implicit biases, which can affect the impartiality of their judgment and pose a threat to the right to a fair trial. In recent years, artificial intelligence (AI) has been increasingly used for a variety of applications in the public domain, often with the promise of being more accurate and objective than biased human decision-makers. Given this backdrop, this research article identifies how AI is being deployed by courts, mainly as decision-support tools for judges. It assesses the potential and limitations of these tools, focusing on their use for risk assessment. Further, the article shows how AI can be used as a debiasing tool, i.e., to detect patterns of bias in judicial decisions, allowing for corrective measures to be taken. Finally, it assesses the mechanisms and benefits of such use.

Keywords • judicial decision-making, judicial biases, artificial intelligence, risk assessment, debiasing

Introduction

Several cognitive and social psychology studies suggest that judges are susceptible to various implicit biases which, unlike overt prejudice, they tend to be unaware of. These can influence their decisions in ways that are problematic considering the duty of impartiality and the right to a fair trial. The desire to increase objectivity, accuracy, and consistency in judicial decision-making has prompted the adoption of artificial intelligence (AI) to assist with different decision points throughout proceedings. At the same time, recognizing judges’ susceptibility to biases raises the issue of how to mitigate them, and it is worth questioning whether AI might have a role to play in it. Hence, the goal of this article is twofold: First, it will describe how AI has been adopted for decision-support in judicial systems, and the challenges aris-
ing from such use; and second, it will evaluate the possibility of using AI for helping to detect and counteract judges’ implicit biases, recommending which extralegal factors should be considered when doing so. It is a theoretical and bibliographic research, drawing on direct and indirect sources for a comprehensive analysis of the theme. The article will proceed as follows: I will first provide an overview of how implicit biases affect judicial decision-making, consequently giving rise to arguments favoring the adoption of algorithms to promote more accurate and objective decisions. Subsequently, I will explore their current use in judicial settings, focusing on decision-support tools that are adopted to promote risk assessments. The implications of using automated procedures in the judicial process will be evaluated, with the desirability of their adoption being called into question. I will then offer a different possibility of AI use in the judiciary, namely, to help identify and counteract judicial bias, therefore increasing fairness and legal certainty, before drawing some conclusions.

Biases in adjudication

There is ample scientific evidence demonstrating how judges – like jurors and laypeople – are prone to both cognitive and social biases. While the former entails some broadly erroneous form of reasoning, the latter entails reasoning based on stereotypes (Zenker 2021). Biases have the potential to reduce the accuracy of a judgment and, throughout different stages of proceedings, can influence judicial decisions. To give some examples of relevant findings:

- Judges’ sentencing decisions and compensation awards were found not only to be anchored by the initial demand made by the prosecutor, but also by random and unrelated factors to the decision at hand (Bystranowski et al. 2021).
- In a criminal investigation scenario, irrelevant contextual information affected judges’ conviction rate, and confirmation bias led them to prefer incriminating investigations (Rassin 2020). Similarly, the pretrial detention of defendants later influenced judges’ assessments of their guilt in criminal cases (Lidén et al. 2019).
- Judges’ decisions were biased by the gender of the parties in studies involving hypothetical cases about child custody and relocation, employment discrimination, and criminal sentencing (Miller 2019; Rachlinski and Wistrich 2021).
- Data analysis of judges’ bail decisions revealed racial bias against black defendants, even after controlling for variables such as criminal history and past pretrial misconduct (Arnold et al. 2020). In a virtual reality courtroom, minority defendants were treated more harshly by evaluators – including judges – during conviction (Bielen et al. 2021).

The idea that judicial decision-making can be influenced by extralegal factors is problematic considering judges’ duty of impartiality and the right to a fair trial. Article 6 of the European Convention on Human Rights (ECHR) establishes the right to a fair trial by an independent and impartial tribunal. Impartiality requires that judicial decisions are based on the objective circumstances of the case, in accordance with the law, and free from external influences. Moreover, it excludes the existence of a prior disposition of the judge’s mind that could lead them to favor or harm either party. The European Court of Human Rights (ECHR) has distinguished between an objective aspect of this requirement, linked with the appearance of impartiality, and a subjective one, linked to “the personal conviction and behavior of a particular judge, that is, whether the judge held any personal prejudice or bias in a given case”. The existence of a subjective approach may lead one to believe that there is an effective remedy to fight against judges’ implicit biases, but such remedy is truly limited. The ECHR has recognized the hardship of establishing a breach of Article 6 on account of subjective partiality, given the difficulty to procure evidence with which to rebut the presumption of impartiality, and has thus justified its common recourse to the objective analysis.

One way to do so is through the implementation of debiasing techniques, which seek to address biases’ negative effects by improving either the decision-making process or some relevant characteristics of the decision-maker (Zenker 2021). Another possibility relates to the adoption of artificial intelligence in judicial systems as decision-support or decision-making tools. ‘Artificial intelligence’ is used as an umbrella term to describe various human-designed technologies that exhibit intelligent behavior, analyzing their environment, and taking actions – with a certain level of autonomy – to achieve specific goals. The use of AI brings with it the promise of more accuracy, objectivity, and consistency, with governments increasingly adopting the technology “to attain greater accuracy when making predictions, replace biased human decisions with ‘objective’ automated ones, and promote more consistent decision-making” (Green 2022, 1 ECHR, Micallef v. Malta, Judgment of 15 October 2009, Application No. 17056/06, p. 22.


Given the high stakes involved in judicial decision-making, the issue of how to mitigate judicial bias is important.
p. 3). However, and at least for the time being, not only do these systems also have several limitations that can further deepen the problem of bias in adjudication, but there is also a risk-magnifying potential associated with AI that is not present with human decision-making (Dietterich 2019). In the following session, I will examine how AI has been adopted in judicial systems, specifically as decision-aid tools for assessing risk, and the challenges posed by this use.

**AI in judicial systems**

To assist adjudication, several countries are experimenting with and integrating digital technologies, particularly AI, in their judicial systems. Applications like advanced case-law search engines, online dispute resolution, or document categorization and screening can potentially lower the cost of dispute resolution and help courts address their backlog of cases, many of which are low-volume, low-value, and low-complexity matters (Steponenaita and Valcke 2020). Furthermore, some evidence suggests that algorithms are better at making policy-relevant predictions than public servants (Kleinberg et al. 2018). This makes the prospects of adopting digital technologies in judicial systems, particularly supportive and advisory AI-based tools, quite promising. On the other side, the use of algorithms to make consequential decisions about the application of public policy to individuals in street-level bureaucracies like courts, police departments, and welfare agencies has been highly controversial (Angwin et al. 2016; Heaven 2020; Allhutter et al. 2020).

Considering this scenario, institutions such as the European Union (EU) and the Council of Europe (CoE) are working towards ensuring that the development, implementation, and use of AI is done in an ethical and lawful way, especially in contexts where there is a high impact on individuals’ fundamental rights. While the EU is in the final stages of approving a regulation creating standardized rules for AI (European Commission 2021), the CoE’s Commission for the Efficiency of Justice adopted the first European Ethical Charter on the use of AI in judicial systems (CEPEJ 2018). In it, the Commission, which is responsible for evaluating European judicial systems and defining concrete ways to improve their performance, provides an ethical framework to guide private and public stakeholders throughout the development and implementation of AI in the judiciary. Continuing this work, in April 2023 CEPEJ also launched a Resource Centre on Cyberjustice and AI, which aims to serve as a publicly accessible focal point for reliable information on AI systems applied in the transformation of judicial systems (CEPEJ 2023). One of its first endeavors was to obtain an overview of these systems, providing a starting point for further examination of their risks and benefits for professionals and end-users. A total of 58 systems were identified in CoE member states, and then classified according to their main field of application. Categories include, e.g., ‘anonymization tools’, which are used for removing identifying information of court users, and ‘natural language processing tools’, used for speech recognition and the automatic transcription of court procedures. My analysis here will focus, however, on the category of ‘decision-support and decision-making’, which encompasses tools meant to facilitate or fully automate decision-making processes in justice systems, considering that some think that highly accurate AI systems could improve the performance of judges, or even come to replace them (Chatziathanasiou 2022).

First, it is worth highlighting that the use of the tools mapped by CEPEJ by judges themselves is still quite limited, and the initiative for their development remains primarily within the private sector, focusing on insurance companies, lawyers, and legal services wishing to reduce the uncertainty and unpredictability of judicial decisions. The French application Predictice, for example, is a predictive justice tool developed to calculate the chances of success of a legal action according to different variables, using jurisprudence analysis algorithms. More recently, it has launched a generative AI tool called Assistant, developed to answer legal professionals’ questions by citing reliable and up-to-date sources (Larret-Chahine 2023). Even though most applications of this kind have their use currently restricted to private agents, “public decision-makers are beginning to be increasingly solicited by a private sector wishing to see these tools […] integrated into public policies” (CEPEJ 2018, p. 14).

Second, not all the applications listed at the Resource Centre can technically be categorized as AI, as is the case for many of the risk assessment tools, mainly used for assessing the risk of recidivism. These make up for the majority of ‘decision-support and decision-making’ tools that have been listed by CEPEJ as being currently used in the public sector, namely by judges. Examples include OASys, the Offender Assessment System used by the prison and probation services in England and Wales (Justice Data Lab 2016), RITA, the Finish Risk and Needs Assessment Form (Salo et al. 2019), or RISC, the Recidivism Assessment Scale adopted in the Netherlands (van Essen et al. n.d.). Risk assessments are perceived and often marketed as an objective means of overcoming human bias in decision-making and have been adopted to assist with several decision points throughout the criminal justice system, from pretrial release to post-conviction sentencing, probation, and parole. These tools do not use new statistical methods commonly associated with AI, such as machine learning (ML), but are rather overwhelmingly based on regression models (Barabas et al. 2018). The main goal of regression is to identify a set of variables (e.g., prior arrest) that are predictive of a given outcome variable (e.g., risk of reoffending). This process can be automated and improved using ML methods (Ghasemi et al. 2021), but still constitute an incremental innovation in the way risk assessments have historically worked, instead of being truly transformational.

One example of a risk assessment tool that incorporates a machine learning approach is the Correctional Offender Management Profiling for Alternative Sanctions, or COMPAS, used
by some United States’ courts to assess the likelihood of recidivism (van Dijck 2022). Ever since an exposé by the news outlet ProPublica revealed that the software was biased against blacks (Angwin et al. 2016), it has become the primary example of the risks posed by algorithmic crime prediction overall. The controversy surrounding its use revolves around what the models measure and intend to measure, the accuracy of the predictions, and whether they might increase inequality and discrimination or otherwise compromise fairness (Mayson 2019; Rudin et al. 2020). The question of (un)fairness is of particular concern, given that risk assessment tools can lead to discriminatory outcomes based on race or ethnicity (Jordan and Bowman 2022).

Furthermore, since the software is proprietary, the data and algorithms are not transparent, neither for the suspect nor for the judge (van Dijck 2022), a problem that was addressed in the case of Loomis v. Wisconsin3 (2016). In this case, while admitting the system’s flaws, the Court claimed that it is up to judges to exercise discretion when assessing a risk score.

In high-stakes decisions such as criminal justice risk assessments, it is common to place emphasis on the decision-makers’ discretion in incorporating algorithmic advice into their decisions, to make their use acceptable even in light of flaws. However, human discretion does not necessarily improve outcomes. Decision-makers are susceptible to automation bias, a tendency to defer to automated systems, reducing the amount of independent scrutiny exhibited when deciding (Parasuraman and Manzey 2010). Similar issues arise when humans collaborate with predictive algorithms. Recent research has found that people are bad at judging the quality of algorithmic outputs and determining whether and how to override those outputs (Green 2022). In simulated pretrial and sentencing decisions, for instance, risk assessments made participants – including judges – place a greater emphasis on its results than on other relevant factors (Green and Chen 2021). It is thus likely that, instead of improving the issue of bias by promoting an ‘objective’ score, the incorporation of results into a decision may nonetheless result in biased outcomes.

The challenges discussed in this section do not necessarily entail a categorical rejection of employing AI for assisting in judicial decisions, but it does raise the question of which uses might be advantageous without presenting a risk to the fairness of a trial. Here, one possibility relates to its use for triaging, allocation, and workflow automation, facilitating some activities that are the main extralegal factors that influence judges when deciding. Based on an examination of the literature on social biases, initial contenders include race, gender, and ethnicity of the defendant and of the judge, the latter for the assessment of ingroup favoritism. And based on findings on cognitive biases, the number of (un)favorable previous decisions by the court, the comparison of caseloads between judges, and whether it is a spe-

---

3 Wisconsin Supreme Court, Wisconsin v. Loomis 2016 WI 68.
cialized court (all being indicative of contrast effects) are variables likely to influence the way judges decide. Other factors besides biases include the time of day at which a decision is made (Shroff and Vamvourellis 2022; Danziger et al. 2011), and temperature (Chen and Loecher 2019; Heyes and Saberian 2019). These are of course only initial suggestions and do not fully encompass all the relevant covariates that can affect a decision. But the advantage of using machine learning techniques for this purpose is precisely that any sort of data can be used to feed the model, enabling patterns and trends to emerge without necessarily requiring a theoretical explanation for such.

In the in-depth study on the use of AI in judicial systems that accompanies CEPEJ’s ethical charter (2018), the commission analyzes the benefits and risks of different applications, encouraging their use to various degrees. While specific judge profiling is highly discouraged, among uses to be considered following additional studies is offering judges an assessment of their activities with an informative aim of assisting in decision-making. Indeed, offering judges’ feedback regarding their decisions is a fundamental step in debiasing, alongside other interventions such as the promotion of general bias awareness, training in rules and representations, exposure to stereotype-incongruent models, and the adoption of scripts and checklists (Wistrich and Rachlinski 2017) – all of which can be directed once bias is identified. Furthermore, AI offers a mechanism of detecting bias in real time, and could hence alert judges to situations where biases are likely to occur (e.g., after a string of positive decisions), allowing them to intervene before a biased decision takes place.

Concluding remarks

The adoption of digital technologies like artificial intelligence in judicial settings often comes from a desire to increase objectivity, accuracy, and consistency in decision-making, improving the quality of decisions traditionally made by humans. However, we have seen how the use of AI for decision-support in adjudication, albeit still not prevalent in CoE member states, can worsen issues already identified in the risk assessment tools that they seek to automate. These include the accuracy (or lack thereof) of their predictions, the reproduction of existing patterns of prejudice and bias, the lack of transparency and opportunities for defendants to challenge their outcomes, and the difficulty of decision-makers to properly evaluate assessments. Thus, instead of using AI to make decisions that traditionally pertain to judges (e.g., assessing the risk of recidivism), a different possibility was offered, namely, to employ the technology for debiasing purposes. AI can help identify the situations where judicial bias is likely to take place, based on the analysis of covariates that, despite being legally irrelevant, have been shown to influence judicial decisions, some of which were listed here. By identifying the instances in which bias commonly arises, it is possible not only to alert judges but also to target debiasing interventions, such as educating judges on the subject and offering feedback on their work, with the ultimate goal of ensuring objectivity and impartiality in their decisions.

Funding • This work received no external funding

Competing interests • The author declares no competing interests.

References


intelligence (Artificial Intelligence Act) and amending certain union legisla-
europa.eu/legal-content/EN/TXT/?uri=celex:52021PC0206, last accessed on
22.01.2024.
CEPEJ – European Commission for the Efficiency of Justice (2018): European
erthic charter on the use of artificial intelligence in judicial systems and
their environment. Strasbourg: Council of Europe. Available online at https://
rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c,
last accessed on 22.01.2024.
CEPEJ (2023): Resource centre on cyberjustice and AI. Available online at https://
public.tableau.com/app/profile/cepej/viz/ResourceCentreCyberjusticeandAI/
AITOOLSINITIATIVESREPORT, last accessed on 22.01.2024.
Ghasemi, Mehdi; Anvari, Daniel, Atapour, Mahshid, Wormith, Stephen, Stockdale,
Keira, Sperli, Raymond (2021): The application of machine learning to a
general risk-need assessment instrument in the prediction of criminal
doi.org/10.1177/0093854820969753
Green, Ben, Chen, Yiling (2021): Algorithmic risk assessments can alter human de-
cision-making processes in high-stakes government contexts. In: Proceedings
of the ACM on Human-Computer Interaction 5 (CSCW). New York, NY: Associ-
ation for Computing Machinery, pp. 1–33. https://doi.org/10.1145/3479562
Green, Ben (2022): The flaws of policies requiring human oversight of govern-
go/10.1016/j.clsr.2021.05681
Heaven, Will (2020): Predictive policing algorithms are racist. In: MIT Technol-
com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-
machine-learning-bias-criminal-justice/, last accessed on 10.01.2024.
app.20170223
Kleinger, Jon; Lakkaraju, Himabindu; Leskovec, Jure; Ludwig, Jens; Mullainathan,
Larret-Chahine, Louis (2023): Predictice lance assistant, une IA générative pour les professionnels du droit. In: Predictice Blog, 26.05.2023 available online at
https://blog.predictice.com/assistant-ia-pour-les-professionnels-du-droit,
last accessed on 10.01.2024.
Jordan, Kareem, Bowman, Rachel (2022): Interacting race/ethnicity and legal
factors on sentencing decisions. A test of the liberation hypothesis. In:
Lidén, Moa; Grans, Mimna; Justin, Peter (2019): `Guilty, no doubt’. Detention pro-
voking confirmation bias in judges’ gsat assessments and debiasing tech-
10683161.2018.1511790
bias-in-bias-out, last accessed on 10.01.2024.
Miller, Andrea (2019): Expertise fails to attenuate gendered biases in judicial
decision making. In: Social Psychological and Personality Science 10 (2),
Parasuraman, Raja; Manzey, Dietrich (2010): Complacency and bias in human
10.1177/0018720810367055
Rachlinski, Jeffrey; Wistrich, Andrew (2021): Benevolent sexism in judges. In:
sandiego.edu/sdlr/vol58/s1/jr313, last accessed on 22.01.2024.
Rassin, Eric (2020): Context effect and confirmation bias in criminal fact finding.
In: Legal and Criminological Psychology 25 (2), pp. 80–89. https://doi.org/
10.1111/lcp12172
Saló, Benny, Laaksonen, Toni, Santtila, Pekka (2019): Predictive power of
dynamic (vs. static) risk factors in the Finnish risk and needs assessment
10.1177/00938548198448793
Steponenaite, Vilde, Valcke, Peggy (2020): Judicial analytics on trial. An assess-
ment of legal analytics in judicial systems in light of the right to a fair trial.
In: Maastricht Journal of European and Comparative Law 27 (6), pp. 759–773.
https://doi.org/10.1177/1023265320981472
Shroff, Ravi, Vamvourellis, Konstantinos (2022): Petriil release judgments and
decision fatigue. In: Judgment and Decision Making 17 (6), pp. 1176–120.
https://doi.org/10.1177/1555256520993098
Rudin, Cynthia; Wang, Caroline, Coker, Beau (2020): The age of secrecy and
unfairness in recidivism prediction. In: Harvard Data Science Review 2 (1),
pp. 1–53 https://doi/10.1162/10990297500009384
SWR (2022): OLG Stuttgart setzt Ki bei Diesel-Klagen ein. In: SWR Aktuell,
baden-wuerttemberg/stuttgart/olg-stuttgart-mit-ki-gegen-flut-von-
dieselklagen-100.html, last accessed on 10.01.2024.
Van Essen, Laurus; Van Alphen, Huib; Van Tuinen, Jan-Maarten (n.d.): Risk assess-
ment the Dutch way. A scalable, easy to use tool for probation reports. In:
cep-probation.org/risk-assessment-the-dutch-way-a-scalable-easy-to-use-
tool-for-probation-reports/, last accessed on 22.01.2024.
Van Dijk, Gijs (2022): Predicting recidivism risk meets AI act. In: European Journal
105610–022–09516–8
Wistrich, Andrew, Rachlinski, Jeffrey (2017): Implicit bias in judicial decision
making. How it affects judgement and what judges can do about it. In: Sarah
Redfield (ed.): Enhancing justice: Reducing bias, pp. 87–130. https://doi.org/
10.31228/osf.io/sz5ma
Zenker, Frank (2021): De-biasing legal factfinders. In: Christian Dahlman, Alex
Stein and Giovanni Tuzet (eds.): Philosophical foundations of evidence
oso/9780198859307.003.0027

GIOVANA LOPES
is a doctoral candidate in “Law, Science and Technology” at the University of Bologna and KU
Leuven and an affiliated research fellow at the Centre for IT & IP Law (CiTiP).
Borderline decisions?:
Lack of justification for automatic deception detection at EU borders

Daniel Minkin*1, Lou Therese Brandner2

Abstract • Between 2016 and 2019, the European Union funded the development and testing of a system called “iBorderCtrl”, which aims to help detect illegal migration. Part of iBorderCtrl is an automatic deception detection system (ADDS): Using artificial intelligence, ADDS is designed to calculate the probability of deception by analyzing subtle facial expressions to support the decision-making of border guards. This text explains the operating principle of ADDS and its theoretical foundations. Against this background, possible deficits in the justification of the use of this system are pointed out. Finally, based on empirical findings, potential societal ramifications of an unjustified use of ADDS are discussed.

Introduction

The potential of artificial intelligence (AI) to revolutionize border management has been recognized for some time (Beduschi 2020). In the European Union (EU), different AI-based technologies for border control are either already in use or are being tested for future deployment, such as biometric identification and verification, risk assessment, or emotion detection (Dumbrava 2021). This contribution focuses on a subset of the latter: The so-called automatic deception detection system (ADDS), part of the iBorderCtrl project funded by the EU, was developed to detect cases of illegal border crossing by video-interviewing travelers to analyze their facial microexpressions for indicators of deceit. The technology is intended to support border guards in their decision-making process, providing recommendations in the form of risk assessments regarding individual travelers. Who gets to cross European borders, an already complex social issue with major implications for migrants and society at large, thus becomes embedded in discourses around AI-based decision support. These discourses pertain not only to one state or one scientific discipline. A responsible use of AI-based systems at border crossing points requires responsible policy-making based on an interdisciplinary and transnational perspective.

Keywords • automatic deception detection, machine learning, emotion recognition, border control, trust

This article is part of the Special topic “AI for decision support: What are possible futures, social impacts, regulatory options, ethical conundrums and agency constellations?,” edited by D. Schneider and K. Weber. https://doi.org/10.14512/tatup.33.1.08

© 2024 by the authors; licensee oekom. This Open Access article is licensed under a Creative Commons Attribution 4.0 International License (CC BY). https://doi.org/10.14512/tatup.33.1.34

Received: 22.08.2023; revised version accepted: 03.01.2024; published online: 15.03.2024 (peer review)
Against this background, in this paper, we bring together epistemological, technological, and social science arguments in order to contribute to an informed assessment of the technology.

After giving a more in-depth description of ADDS, its purpose and theoretical basis, we focus on two interlinked questions:

1. What are the concerns about the use of ADDS and are they warranted? We examine epistemological and empirical arguments against the deployment of ADDS. The first group of arguments targets the system’s theoretical foundation, contending that ADDS rests on a scientifically unfounded basis. The second group attempts to show that the mechanisms underlying ADDS are not sufficiently accurate.

2. Having discussed the above concerns, we turn to the second question: Given the concerns outlined, what would be the social implications of using ADDS in terms of public trust? The concept of public trust is related to various dimensions of technology assessment (TA) such as trustworthy technology, acceptance of new technologies, as well as the promotion and maintenance of trust in technology-related policy-making (Weydner-Volkmann 2021). We address these aspects at the end of the paper, arguing in favor of more transparency in the implementation of systems such as ADDS.

The automatic deception detection system

Purpose and working principle
ADDS is a machine learning (ML) based system designed to identify deception when crossing a state border (O’Shea et al. 2018; Podoletz 2023). It has been dubbed an “AI Polygraph” or “AI lie detector” (Kaminski 2019, p. 178). ADDS is part of iBorderCtrl, which is designed to facilitate and accelerate the registration and control of travelers coming to the EU, including refugees. It has already been tested in Hungary, Latvia, and Greece, with the test phase ending in 2019. Currently, it is not known whether and in what form iBorderCtrl will be deployed; however, with the EU extensively testing this kind of technology and funding several related border control and surveillance projects (iBorderCtrl 2023), critically examining these systems remains relevant.

iBorderCtrl works in the two stages of pre-registration and border crossing. It includes procedures such as biometric identification, document matching, risk analysis, and ADDS, on which this contribution focuses. During pre-registration, travelers undergo an online video interview with a police avatar. ADDS analyzes the recorded interviews, more specifically combinations of the travelers’ microexpressions, very subtle facial expressions, normally invisible to the naked eye, to quantify the probability of deceit. O’Shea et al. (2018, p. 3) point out a difference between microgestures and microexpressions: The former are “more fine-grained and require no functional psychological model of why the behaviour has taken place”. However, since the original description by Rothwell et al. (2006, p. 759) uses the term “microexpression”, we follow this publication. Based on the microexpression analysis and other components of the iBorderCtrl system, a risk estimation regarding the traveler is provided. The system is intended as a human-in-the-loop system: When travelers attempt to cross the border, a border guard makes the final decision after performing a security check against the background of the data provided by the system.

The main component of ADDS is a subsystem called “Silent Talker” (ST) (iBorderCtrl 2018, p. 15). By using several artificial neural networks, ST learns to recognize combinations of microexpressions (by means of supervised or unsupervised methods). The actual classification of these combinations as truthful or deceptive is based on a conceptual model of non-verbal behavior (NVB): “This model assumes that certain mental states associated with deceptive behavior will drive an interviewee’s NVB when deceiving. These include Stress or Anxiety (factors in psychological Arousal), Cognitive Load, Behavioral Control and Duping Delight” (O’Shea 2018, pp. 3–4). It is worth noting that a combination’s classification as deceptive cannot be regarded as proof of deception but as a probabilistic result obtained by inductive learning.

The assumption that microexpressions are capable of revealing deceptive intentions thus forms the theoretical basis of ST, which will be further discussed in more detail in the next chapter.

Theoretical foundation
Deception consists of or involves mental states, especially intentions, while microexpressions are a kind of behavior open to intersubjective investigation. ST and ADDS analyze the latter and thereby provide a basis for human actors to obtain information about the former. Against this backdrop, the question arises as to which possible combinations of microexpressions can indicate an intention to deceive on the part of the subject. In other words, the development of deception detection systems requires a psychological theory about the connections between microexpressions and mental states. In their description of the ST, the inventors explicitly state that they take some key elements from the psychological work of Paul Ekman, such as Ekman’s definition of microexpression (Rothwell et al. 2006, p. 759).

Ekman became famous for his cross-cultural studies of emotional expression. In the 1960s, for instance, he asked indigenous people in New Guinea, at that time largely isolated from the Western world, to assign terms like ‘sad’ to pictures of Europeans expressing different emotions. Experiments with various cultures as well as later studies (Elfenbein and Ambady...
2002) resulted in relatively high accuracy rates, leading to the assumption of universal emotional expression. Regarding universal microexpressions, Ekman analyzed video recordings of proven liars frame by frame, finding that subjects with deceptive intentions could not consciously control all facial muscle movements while experiencing a particular emotion (Ekman 1985, p. 133).

Ekman’s theory indicates, therefore, that some unconscious and uncontrollable facial expressions can provide evidence of deception. However, this idea is controversial; in the coming chapter, we will explore criticisms of microexpression analysis and ADDS.

Criticism

Deception detection has been criticized as ‘pseudoscience’ (Whittaker et al. 2018). Against the background of such normative characterizations, we want to assess if the use of ADDS is justified, focusing on the system’s theoretical background and its accuracy; the first aspect is the subject of an epistemological criticism, the second of an empirical one.

Epistemological criticism

The psychological foundation of ST and ADDS has been criticized as flawed, which, according to the systems’ opponents, means their use cannot be justified. A major part of this criticism is an observed lack of scientific consensus on the assumption that deceptive intentions can be derived from microexpressions. In general terms, the epistemological criticism states that the use of deception detection systems is not justified unless there is widespread agreement on their theoretical foundations (Podoletz 2023, p. 1071). Given that Ekman assumes one can derive deceptive intentions from microexpressions, it is reasonably only justified if their accuracy rates exceed those of human experts. Since human actors are expected to make the final decision, the systems would otherwise have no added value; in the worst case, they would direct human experts towards wrong decisions due to the perceived authority of automated outputs (Helm and Hagendorff 2021).

While this line of thought appears convincing, there are some limitations. To start with, although it is unclear how much disagreement is too much, it seems uncontroversial that perfect consensus is not necessary for justifying the use of a system. On the other hand, even if there were a perfect consensus on the theory as a basis of a system, the use of this system would not be justified without a sufficiently high level of accuracy. This indicates that a lack of consensus on a system’s theoretical basis does not necessarily impact the justification of its use; it can be argued that it is irrelevant whether ST is based on a controversial theoretical basis as long as it is able to distinguish deceptive statements from truthful ones with sufficient accuracy. In the third part of this paper, we will revisit the relevance of the theoretical foundation.

Empirical criticism

The empirical criticism of ADDS considers the accuracy of such systems (Sánchez-Monedero and Dencik 2022). As part of iBorderCtrl, ADDS is intended to support the decision-making of border guards. The use of such subsidiary systems, particularly in high-risk application contexts like border control, is arguable only if the accuracy rates exceed those of human experts. Since human actors are expected to make the final decision, the systems would otherwise have no added value; in the worst case, they would direct human experts towards wrong decisions due to the perceived authority of automated outputs (Helm and Hagendorff 2021).

The concern is that in the case of ST this condition is not met: Its accuracy rate has been reported to be 63 to 70% (Rothwell et al. 2006) and 74.6% (O’Shea et al. 2018). Empirical find-
ings suggest that while human performance is below these values for untrained subjects, ST’s performance does not exceed human experts. According to a meta-analysis, subjects without special training perform only slightly better than chance (54%) when attempting to distinguish deceptive from truthful statements (Bond and DePaulo 2006). However, the mean accuracy of trained parole officers has been found to be 76.7% (Porter et al. 2000) and thus beyond the highest values found for ST.

A second empirical problem is that ST was trained and tested on a surprisingly small number of subjects under controlled experimental conditions. Rothwell et al. worked with 39 subjects and O’Shea et al. with 30, with most of them being male Europeans. Since the aim of this system is to control non-EU citizens, Rothwell et al. openly admit this could lead to data bias (Rothwell 2006, p. 768): Underrepresenting certain groups – such as people of color or women – in AI training datasets can lead to unreliable assessments regarding individuals belonging to these groups and therefore to discriminatory outcomes (Brandner and Hirschbrunner 2023; Selbst 2017). The results of the real-life tests in Hungary, Latvia, and Greece have not been publicly disclosed, and it is thus so far unclear whether they show biases.

These concerns are independent of the epistemological criticism since they relate solely to the accuracy of the system in question. Thus, even ignoring potential flaws in the theoretical foundations, there seems to be no sufficient justification for deploying ADDS. Taking into account the addressed empirical criticism, the next chapter assesses the potential social implications of ADDS. To do so, we build on empirical studies on public attitudes toward the police and AI-based technology.

Social implications

Trust in automated deception detection

AI-based policing and border control evoke divergent public attitudes. For controversial real-time facial recognition technology, acceptance appears to depend on the general trust invested in the police, with a nearly fifty/fifty split regarding the question if the police should be able to use this technology (Bradford et al. 2020). While trust is a much debated, complex notion, someone trusting in an institution like the police can be described as “having confidence that the institution is reliable, observes rules and regulations, works well, and serves the general interest” (Devos et al. 2002, p. 484). Trust in the police varies greatly between European nations and regions, particularly between Scandinavia (high trust) and Eastern Europe (low trust) (Pfister 2020).

Public attitudes toward AI decision-making and support are far from uniform and depend on multiple interrelated factors, such as application context, geographical and cultural differences, or other (perceived) characteristics of the respective technology, such as fairness and transparency (Starke et al. 2022). It has been found that in the application context of justice, automated decisions are perceived as less risky and fairer than decisions made by human experts (Araujo et al. 2020). Others, however, suggest a preference for decisions made by human actors in a policing context (Hobson et al. 2023). Human decisions to accept or reject AI suggestions are furthermore not only contingent on the confidence vested in the system but also on personal self-confidence (Chong et al. 2022). Both over- and undertrust can thus lead to problematic outcomes in high-risk situations such as border control, where humans are meant to make reliable final decisions based on AI suggestions. Given the lack of social consensus, the coming paragraphs assess the implications of both low and high trust in a system like ADDS.

Implications of low trust

Public trust is fundamental for the societal acceptance of AI-based systems and therefore their sustainable adoption (Gil-lespie et al. 2023). The deployment of unreliable AI-based technology can actively lead to a loss of trust in public institutions (Starke et al. 2022). While, to our knowledge, trust in automated deception detection in a border control context is yet to be studied empirically, automated emotion recognition appears to predominantly evoke negative attitudes: Interviewees describe it as “invasive” and “scary” (Andalibi and Buss 2020, p. 6). Those who believe emotion detection to be accurate can also perceive this accuracy as concerning, i.e., as a threat to human agency (Grill and Andalibi 2022). Others question if the technology can work at all due to the complexity of human emotions. This indicates that academic criticisms of the theoretical basis and empirical accuracy of ADDS are also reflected in citizen concerns.

While ADDS is meant to control non-EU nationals and therefore poses no immediate personal risk to EU citizens, based on the qualitative studies just mentioned it seems plausible to assume that many would fundamentally oppose the use of deception detection in a high-risk setting such as border control. Particularly persons with a general “anti-surveillance” viewpoint (Ezzeddine et al. 2023, p. 869) emphasize the importance of personal freedom over security and oppose all police AI that might flag individuals as suspicious, regardless of who the system is used on. Those generally critical and distrustful of AI policing might perceive the use of ADDS at European borders as threatening to human agency rather than as a trustworthy security measure. Given the already uneven trust in the police within the EU, deploying ADDS might further erode this shaky ground, leading to increased dissonance between nations and political unrest.

The perceived transparency of systems impacts trust, with higher transparency entailing more trust (Aysolmaz et al. 2023). As has been shown in TA studies, trust-building communication cannot only consist of conveying technical aspects such as reliability (Weydner-Volkmann 2021). Meaningful transparency should also include justifying the logic, reasoning, and legality of (semi-)automated decisions (Malgieri 2020). In the case of ADDS, this would inevitably include disclosing and explaining the technology’s contentious theoretical basis and potentially
Implications of high trust

Freedom of movement of EU citizens is one of the cornerstones of the European project. Yet, with the Russian invasion of Ukraine, high immigration and growing support for far-right parties, even internal borders such as Germany-Poland undergo more rigorous checks, while the majority of EU citizens support stricter external border protection and up to a third think individual nations should control their own borders (BrusselsReport.eu 2022). As opposed to groups who fundamentally oppose AI policing, parts of the population passively trust any police action on principle (Bradford et al. 2020), which might lead them to be more accepting of technologies they would oppose in other contexts. Given that the iBorderCtrl system is meant to control non-European migrants while EU citizens maintain privacy, a “Not Me group” (Ezzeddine et al. 2023, p. 872) might also be prevalent in this case; these individuals endorse AI policing for their personal safety but not on their own data and might therefore trust in systems like ADDS.

Comprehension is not a necessary prerequisite for trust. Instead, people often trust in things they find too complex to understand (Reinhardt 2023). Considering its justification issues, trust in lie detection technology might be misplaced since unreliable systems can lead to biased decisions; it has for instance been found that emotion detection systems can have racist bias (Rhue 2018). Not only underrepresenting populations can lead to discriminatory outcomes but also overrepresenting already marginalized groups (Bacchini and Lorusso 2019). If predominantly non-EU citizens’ data are fed into ADDS, the system might for instance learn to disproportionately classify the micro-expressions of individuals of non-European descent as deceptive.

The use of ADDS for border control might not only continue inequalities and discriminatory dynamics but, by automating them, embed them further into the social fabric of an already divided and crisis-ridden Europe.

The question of transparency is again relevant here, but in the context of actively fostering distrust (Ammicht Quinn 2015). Distrust can mobilize citizens to refuse using certain technologies or actively protest them (Büscher and Sumpf 2015), which can incentivize governments and industries to more carefully assess the potential harms of deploying systems such as ADDS. In the case of ADDS, EU citizens’ data would not be analyzed; for EU citizens, expressing distrust in the form of refusal is therefore not possible and those with a “Not Me” perspective might overall not be interested in protesting the technology. However, deploying similar systems at EU borders has the potential to perpetuate and aggravate harmful social inequalities and therefore affect all parts of society. EU citizens should thus be comprehensively informed about the described risks and uncertainties in order to facilitate reasonable distrust in – and therefore resistance against – potentially unjustified technology.

Conclusion

Both immigration and AI-based systems are complex and controversial societal issues. With iBorderCtrl, the EU has attempted to find solutions to the former via the latter. This paper has highlighted the importance of justification for the use of such a system, particularly considering public trust. At the current state, we observe a lack of justification on two fundamental levels: The theoretical basis of deception detection is highly contentious on an epistemological level. From an empirical perspective, varying accuracy rates achieved in small-sample studies do not seem sufficient to justify the usefulness of the technology compared to human experts. At the same time, public

Deploying automatic deception detection systems might further erode the shaky ground of trust in the police within the EU, leading to increased dissonance between nations and political unrest.

References


Funding • This article is based on research in the projects “Trust in Information”, funded by the Ministry of Science, Research and Arts Baden-Württemberg, and “PEGASUS”, funded by the German Federal Ministry of Education and Research.

Competing interests • The authors declare no competing interests.


Situativität, Funktionalität und Vertrauen: Ergebnisse einer szenariobasierten Interviewstudie zur Erklärbarkeit von KI in der Medizin

Manuela Marquardt, Philipp Graf, Eva Jansen, Stefan Hillmann, Jan-Niklas Voigt-Antons


Situativity, functionality and trust: Results of a scenario-based interview study on the explainability of AI in medicine

Abstract • A central requirement for the use of artificial intelligence (AI) in medicine is its explainability, i.e., the provision of addressee-oriented information about its functioning. This leads to the question of how socially adequate explainability can be designed. To identify evaluation factors, we interviewed healthcare stakeholders about two scenarios: diagnostics and documentation. The scenarios vary the influence that an AI system has on decision-making through the interaction design and the amount of data processed. We present key evaluation factors for explainability at the interactional and procedural levels. Explainability must not interfere situationally in the doctor-patient conversation and question the professional role. At the same time, explainability functionally legitimizes an AI system as a second opinion and is central to building trust. A virtual embodiment of the AI system is advantageous for language-based explanations.

Keywords • explainability, XAI, AI in healthcare, embodied AI, voice dialog system

Zielsetzung

In den letzten Jahren erfolgte eine stete Zunahme von Systemen künstlicher Intelligenz (KI) im medizinischen und therapeutischen Bereich: von Sprachassistentensystemen zur Pflegedokumentation über Chatbots zur Förderung der mentalen Gesund-


**Methode**


**Abb. 1: Übersicht über die Konstruktion der Szenarien.**

Quelle: eigene Darstellung
Für diese Studie entwickelten wir zwei Baseline-Szenarien mit je drei Fallvignetten. Die Vignetten variieren im Einfluss auf die Entscheidung, die ein KI-System mittels des Interaktionsdesigns sowie des Umfangs der verarbeiteten Daten nimmt (siehe Abb. 1).


Die Rekrutierung erfolgte über professionelle Netzwerke. Die Interviews fanden zwischen dem 12.06.23 und dem 14.08.23 statt und dauerten durchschnittlich 57 Minuten. Für die Auswertung der transkribierten Interviews verwendeten wir die Methode der Grounded Theory (Pentzold et al. 2018).

### Inhaltliche Auswertung


#### Interaktionale Ebene


Die von der KI abgegebene Erklärung oder gar das Aufdecken von „Fehlern“ empfinden die Interviewten als unpassende Zweitmeinung, die die Fachkompetenz infrage stellt und das Vertrauensverhältnis gefährdet.


### Tab. 1: Übersicht über die Teilnehmenden.

<table>
<thead>
<tr>
<th>Nr.</th>
<th>Alter</th>
<th>Ausbildung</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>40</td>
<td>Psychologie</td>
<td>Wissenschaftliche Mitarbeiterin</td>
</tr>
<tr>
<td>2.</td>
<td>70</td>
<td>Zahnmedizin</td>
<td>Oberarzt</td>
</tr>
<tr>
<td>3.</td>
<td>35</td>
<td>FA Dermatologie</td>
<td>Funktionsoberarzt</td>
</tr>
<tr>
<td>4.</td>
<td>50</td>
<td>FA Psychosomatik</td>
<td>Oberarzt</td>
</tr>
<tr>
<td>5.</td>
<td>67</td>
<td>FA Pädiatrie &amp; Neonatologie</td>
<td>Chefarzt</td>
</tr>
<tr>
<td>6.</td>
<td>45</td>
<td>FA Dermatologie</td>
<td>Fachärztin</td>
</tr>
<tr>
<td>7.</td>
<td>73</td>
<td>Orthopädie</td>
<td>Arztassistenz</td>
</tr>
<tr>
<td>8.</td>
<td>35</td>
<td>FA Radiologie</td>
<td>Fachärztin</td>
</tr>
<tr>
<td>9.</td>
<td>57</td>
<td>Krankenpfleger</td>
<td>Hygieneassistentarin</td>
</tr>
<tr>
<td>10.</td>
<td>52</td>
<td>Patient*innenvertretung</td>
<td></td>
</tr>
<tr>
<td>11.</td>
<td>53</td>
<td>Medizin</td>
<td>Gesundheitsmanagerarin</td>
</tr>
<tr>
<td>12.</td>
<td>60</td>
<td>Patient*innenvertretung</td>
<td></td>
</tr>
<tr>
<td>13.</td>
<td>42</td>
<td>Psychosomatik</td>
<td>Psychotherapeutarin</td>
</tr>
<tr>
<td>14.</td>
<td>42</td>
<td>FA Kardiologie</td>
<td>Fachärztin</td>
</tr>
<tr>
<td>15.</td>
<td>38</td>
<td>FA Gynäkologie</td>
<td>Fachärztin</td>
</tr>
<tr>
<td>16.</td>
<td>28</td>
<td>Medizinstudent</td>
<td></td>
</tr>
</tbody>
</table>

Quelle: eigene Darstellung

### Prozessuale Ebene


Situativität


Funktionalität

Der Aspekt der Funktionalität umfasst die Bandbreite von Fähigkeiten und Aufgaben, die einer KI zugewiesen werden können, welche Erwartungen im medizinischen Alltag an sie gerichtet sind und inwiefern Erklärbarkeit hier interferiert. Während


Vertrauen kann durch fachspezifische Standards, etwa im Sinne von „Datenbankstandards“ (10), sowie der Absicherung durch medizinische Fachgesellschaften oder Verbände mittels Zertifikaten externalisiert werden, „ähnlich wie bei einem neuen Medikament“ (5). Von besonderem Interesse für die Befragten war, wer in welchem Umfang welche Daten für das Training der KI berücksichtigt hat (5, 6, 9, 10, 16) und welche Verantwortlichkeit daraus resultiert (9). Sollten sich bestimmte KI-Systeme als „Goldstandard“ (11) durchsetzen, würde dies den Bedarf an Erklärbarkeit des allgemeinen Funktionsrein reduzieren.


Folgerungen
Vertrauen

Ein zentrales Leitmotiv ist Vertrauen, das aufgebaut, in Anspruch genommen, untergraben oder externalisiert wird und mit dem Bedarf an Erklärbarkeit eng verwoben ist. Durch Erklärungen kann das Vertrauen in eine KI gestärkt werden, was den Bedarf an Erklärbarkeit in der wiederholten Nutzung mit der Zeit verringert, wenngleich Kontrolle durch den Zugriff auf die ‘Rohdaten’ gegeben sein sollte. Dieser Befund steht in Einklang mit der Unterscheidung von Vertrauen als Einstellung und ‘reliance’ als Verhalten ‘[…] that follows the level of trust’ (Scharowski et al. 2023, S. 4). Kloker et al. (2022, S. 1) unterstreichen die vermittelnde Rolle von Erklärbarkeit zwischen Vertrauen und Vorsicht: ‘[…] Maximizing trust is not the goal, rather to balance caution and trust to find the level of appropriate trust’. Dies spiegelt sich in unseren Ergebnissen wider: Das Vertrauen in die Sicherheit einer KI geht dialektisch aus dem Umgang des Systems mit Unsicherheit hervor, weil dieser zu einer eigenständigen Reflexionsleistung anregt und einer ‘overreliance’ vorbeugt.


Funding · This article received funding by the German federal ministry of education and research (BMBF) as part of the MIA-PROM project (Multimodal Interactive Assistance for the Collection of Patient Reported Outcome Measures).

Competing interests · The authors declare no competing interests.

Literatur

Jenkins, Nicholas; Bloor, Michael; Fischer, Jan; Berney, Lee; Neale, Joanne (2010): Putting it in context. The use of vignettes in qualitative interviewing. In: Qualitative Research 10 (2), S. 175–196. https://doi.org/10.1177/1468794109356937

Kloker, Anika; Fleiß, Jürgen, Koeth, Christoph, Kloiber, Thomas, Ratheiser, Patrick; Thalmann, Stefan (2022): Caution or trust in AI? How to design XAI in sensitive use cases? In: AMCIS 2022 Proceedings 16.


Powell, John (2019): Trust me, I’m a chatbot. How artificial intelligence in health care fails the Turing Test. In: Journal of Medical Internet Research 21 (10), S. e16222. https://doi.org/10.2196/16222


MANUELA MARQUARDT
ist ausgebildete Techniksoziologin und arbeitet seit 2019 als wissenschaftliche Mitarbeiterin an der Charité zu rehabilitationswissenschaftlichen Themen. Seit 2022 promoviert sie im Projekt MIA-PROM.

PHILIPP GRAF
ist Techniksoziologe und Doktorand an der TU Chemnitz zum Thema sozialer Robotik. Seit 2022 arbeitet er für die Hochschule München im Projekt MIA-PROM.

DR. EVA JANSSEN
ist Medizintechnologin und seit 2021 am Institut für medizinische Soziologie und Rehabilitationswissenschaften der Charité.

DR. STEFAN HILLMANN

PROF. DR. JAN-NIKLAS VOIGT-ANTONS
ist seit 2021 Professor für Angewandte Informatik mit dem Schwerpunkt immersive Medien an der Hochschule Hamm-Lippstadt. Er leitet das Immersive Reality Lab, welches sich auf die Erforschung von immersiven Anwendungen sowie auf Schnittstellen zwischen Menschen und Technik fokussiert.
Abstract • Tools based on machine learning (so-called artificial intelligence, AI) are increasingly being developed to diagnose malignant melanoma in dermatology. This contribution discusses (1) three scenarios for the use of AI in different medical settings, (2) shifts in competencies from dermatologists to non-specialists and empowered patients, (3) regulatory frameworks to ensure safety and effectiveness and their consequences for AI tools, and (4) cognitive dissonance and potential delegation of human decision-making to AI. We conclude that AI systems should not replace human medical expertise but play a supporting role. We identify needs for regulation and provide recommendations for actions to help all (human) actors navigate safely through the choppy waters of this emerging market. Potential dilemmas arise when AI tools provide diagnoses that conflict with human medical expertise. Reconciling these conflicts will be a major challenge.

Künstliche Intelligenz in der Melanom-Diagnose: Drei Szenarien, Kompetenzverschiebungen, Regulierungsbedarf und Umgang mit Dissens zwischen Mensch und KI


Keywords • melanoma, diagnosis, artificial intelligence, patient–doctor relationship, diagnostic accuracy
tion, and dermatoscopy (examination of a skin lesion by using an epiluminescence microscope). In cases where there is suspicion, the mole is surgically removed and sent for histopathological testing to a specialist laboratory to determine malignancy. This standardized procedure with clear diagnostic outcomes in suspicious cases is an ideal dataset for training and utilizing machine learning tools (so-called artificial intelligence, AI). AI in our understanding is a computational approach that uses knowledge gained from training cases to identify patterns and make predictions from input data. Specialized AI for dermatological image recognition has surpassed the performance in identifying skin cancer compared to human dermatologists (Esteva et al. 2017; Pham et al. 2021). In their review of 272 studies Jones et al. (2022) found solid performance of AI systems (89% accuracy; 95% Confidence Intervals: 60%–100%), which suggests that patients and clinicians can expect AI to be an asset for diagnosing melanoma. However, none of the studies included acceptance measures on the part of clinicians or patients demonstrating a limitation in the current approach towards technology assessment of AI in diagnostics. The controlled experimental findings should additionally be validated in medical practice to arrive at a realistic assessment of performance and practical fit. This is particularly relevant for a stringent technology assessment that evaluates AI in diagnostics in practical settings. While technology assessment studies for AI diagnostics exist (Schreier et al. 2020; Schwendicke et al. 2021), their underlying frameworks often times neglect categories specific for AI technologies such as cybersecurity and explainability (Farah et al. 2023).

With this conceptual contribution we aim to shed light on four dimensions of AI employment in melanoma diagnosis, i.e., possible futures, social impacts, regulatory options, and ethical conundrums and agency constellations to inform technology assessment researchers about further areas for assessing AI systems in (melanoma) diagnostics along relative axes of analyses. We aim to provide insights for the following questions, framing our discussions within the German healthcare and regulatory system:

1. Possible futures: What role can AI play in diagnosing melanoma?
2. Social impacts: What competencies and changes in roles follow the comprehensive introduction of AI in diagnostics?
3. Regulatory options: Where is AI located between legal regulations and medical evidence-based guidelines?
4. Ethical conundrums and agency constellations: Who decides regarding diagnosis of melanoma – AI or human?

Methods

We conducted a narrative review of the literature on AI in melanoma diagnosis focusing on the aspects 1) scenarios, 2) social impacts, 3) regulatory options, and 4) ethical facets. Accordingly, we used the following search term to find articles in the databases PubMed and Google Scholar: [: Artificial intelligence OR AI] AND [melanoma] AND [diagnos*] AND [[Scenario OR vision OR future] OR [social] OR [legal OR regulat*] OR [ ethic*] OR [ELSI]]. We removed duplicates, screened the remaining articles, and selected the most fitting ones for the four topics based on our expertise. We also screened reference lists of the selected articles to find additional sources. For the scenarios, we conducted an initial brainstorming to develop scenarios as impulses for possible futures and then consulted further literature for details or contrasts.

Results

Possible futures

AI diagnosis can be applied in different settings by different stakeholders. We will focus on the following three scenarios as they cover the outpatient and inpatient healthcare provision in Germany as well as self-administered care outside professional work: (1) second opinion for a dermatologist in outpatient care, (2) triage and prioritization within a dermatologist clinic, and (3) patient self-monitoring.

Scenario 1: Second opinion in outpatient care

In this scenario, dermatologists upload pictures of suspicious moles to an AI database to obtain a second opinion. The system extracts relevant image features, compares them to a database of expert annotations, and generates a second opinion report highlighting potential areas of concern and providing a diagnosis. As such, the AI system complements the initial human assessment providing an additional layer of confidence reducing diagnostic errors. This scenario follows the interaction mode between clinicians and their machines described by Braun et al. (2021). Such a second opinion system might however transform over time: In a first step, time-restrained dermatologists realize that the AI tools provide (1) dissenting and potentially more reliable diagnoses (2) more efficiently than they can. These attributions could make the AI-based second opinion the first or only opinion. Consequently, dermatological expertise might be removed from the process and tasks regarding diagnosis delegated to non-dermatologists, e.g., to medical technical staff. As a result, dermatological hegemony in diagnosing melanoma is challenged. Companies offering AI tools might target general practitioners providing AI-based dermatological expertise. As such, a system initially providing a second opinion to specialist doctors ultimately might lead to spreading dermatological expertise across disciplines whilst removing specialist doctors from this task.

Scenario 2: Triage in a dermatologist clinic

In this scenario, the AI system extends the process of prioritizing patients in a dermatology clinic. When patients arrive for skin examinations, images of their moles are captured and fed into the AI system. The algorithm compares the images to a database of annotated melanoma cases and provides a risk
assessment score for each mole. Based on this score, the system prioritizes patients, flagging those with higher risks for immediate attention by dermatologists. The specialist doctors can then disregard those moles deemed benign by the AI focusing on the prioritized cases. This generates efficiency gains needed in a strained system: Already today, waiting times for outpatient dermatological appointments last 4.9 weeks with urban-rural variations (Krensel et al. 2015). Until 2035, the number of German regions underserved or without any dermatological specialists are expected to increase by 129% and 700%, respectively, in a forecast with moderate demographic changes (Kis et al. 2017).

Scenario 3: Patient self-monitoring
In this scenario, medical laypeople regularly use an AI smartphone app to monitor their moles instead of screening in outpatient dermatological care. Instructed by the app, users capture standardized images of their moles and upload them to a database. The app compares the images to a database of annotated melanoma cases and provides a risk assessment for each mole together with recommendations for further action. Such a scenario follows the interaction mode between patients and machines (Braun et al. 2021). AI-supported self-monitoring empowers users to participate actively in their skin health and facilitates early detection of melanomas, potentially leading to timely medical intervention and improved health outcomes. However, from patients and doctors alike. Second, AI tools need to perform reliably with sufficient specificity and sensitivity. Third, regulatory assurance must be provided to enable the use and billing of AI tools as medical services. Regulatory guidance would also be the basis for assessing harm caused by the tools’ appraisal systems, i.e., false positive or false negative diagnoses or prioritizing the wrong patients. Finally, it is important to acknowledge that imaging represents only one facet of melanoma differential diagnosis next to anamnestic conversation about progression of the mole’s appearance, itching, and bleeding.

Social impacts
Social impacts of AI technology in melanoma diagnosis vary between individual and societal perceptions of the medical profession. Both competencies of dermatologists and the framework of evidence-based medicine are being scrutinized. Patients might be – or at least feel – empowered and informed about their health. Those impacts have consequences for the configuration of the patient-doctor relationship.

Competencies and expertise
Technology has played a crucial role in diagnosing medical conditions. X-rays, magnetic resonance imaging, and electroencephalography all offer literal insights into the human body and aid in diagnostic procedures across medical disciplines such as orthopedics, neurology, or oncology. These technologies are primarily used as tools expanding the diagnostic repertoire of medical professionals. Within this history, AI serves as a continuation of established procedures integrating technology into diagnostics.

Through AI, the ideal of the ‘informed patient’ might actualize with increased adherence and responsibility for the patient’s own health.

self-empowerment oftentimes coincides with personal responsibility that necessitates the willingness of patients and their acceptance of new technologies. With 9.6 annual consultations per capita, Germany has the second most doctor consultations in the EU (OECD 2023). Thus, the German healthcare system relies heavily on the trust relationship between patients and doctors. Shifting health responsibility from the patient-doctor dyad towards the patient-machine interaction might encounter barriers of acceptance.

Implications
The three possible futures answer in varying constellations current debates in healthcare provision characterized by resource scarcity. AI tools potentially provide efficiency gains by automating processes – diagnosing time-consuming difficult cases (scenario 1), prioritizing patients (scenario 2), or providing initial appraisal for patients (scenario 3).

Three conditions seem to be necessary for successful implementation. First, all scenarios assume acceptance for AI tools.

https://doi.org/10.14512/tatup.33.1.48
nosis, but it shifts the competencies needed – “[h]uman beings are, indeed, always necessary. But literally anyone can do the job, provided he is trained to it. Henceforth, men will be able to act only in virtue of their commonest and lowest nature, and not in virtue of what they possess of superiority and individuality” (Ellul 1964, pp. 92–93).

Evidence-based medicine scrutinized

The inherent complexity of AI tools contributes to their intrigue, as they evoke the notion of fortune telling, a topos deeply ingrained in human imaginaries. Examples of entities with predictive capabilities include the ancient oracles in Delphi or Cichyrs, the prophecy about the chosen one in the Harry Potter series, or the clairvoyant ‘precogs’ in The Minority Report. An AI system capable of providing believable predictions bears similarities to a technical version of these transcendent revelations from mythological narratives. Returning to such narratives stands in contrast to evidence-based medicine that demonstrated its effectiveness by achieving better health outcomes (e.g., taming deadly diseases or raising life expectancy) in an understandable, reproducible way. As such, medical professions are faced with a difficult task to reconcile their success using reproducible, experimental methods with novel technologies outperforming humans in certain confined tasks using inscrutable computational methods. In the light of the success rates in image recognition, individual dermatologists understandably start to question their own expertise.

Patient empowerment

Where medical professionals might struggle with shifts in competencies, patients might welcome such a transformation. With the introduction of AI tools, the scarce resource of specialist medical expertise becomes omnipresent in their pockets. Scenario 3 mentioned above demonstrates how AI tools might enhance patients’ perceived self-efficacy, health literacy, and health outcomes.

However, the success of such a transformation depends on both the system’s accuracy and the users’ expectations. Sensitivity and specificity as indicators for accuracy need to reach high thresholds, and patients’ performance and effort expectations must be met for AI tools to unfurl their potential (Venkatesh 2022). Technical and user mistakes can create an erroneous impression of safety to the detriment of the patient. In that sense, a wrongly applied AI tool resembles an FFP-2 mask covering the mouth but not the nose.

The patient-doctor relationship

With the advent of AI systems, potentially life-threatening diagnoses are presented to patients with little contextualization using incomprehensible methods. Unsettled patients then consult doctors who are tasked with managing information whose origin they cannot reproduce or comprehend. Efforts to make AI analyses explainable could lead to more transparency and understanding for human patients and doctors alike (WHO 2021). Currently, however, opaque analytical processes prevail in AI systems. Findings about the patient-doctor relationship illustrate that healthcare provision is more than the communication of facts (Ridd et al. 2009). Rather, reciprocal trust and empathic communication are relevant vehicles to generate better health outcomes for the patient (Chandra et al. 2018). Indeed, analyses of text responses from doctors compared to AI text generation in online forums indicate better quality and more empathy in the AI responses (Ayers et al. 2023). However, human-human relationships with regard, trust, and empathy might be preferable depending on the cultural context.

Implications

As a social impact, competencies potentially shift in several directions. The competency of diagnosing melanoma based on images might shift from dermatologists to technical staff equipped with AI tools. This process frees dermatologists’ resources for other aspects of differential diagnosis where further AI tools might be utilized (see scenarios 1 and 2 above) or lays off abundant dermatologists. AI could thus extend or replace human dermatological expertise. On another dimension, patients might enhance their health literacy using AI systems (Au et al. 2023). Patients encounter medical professionals at eye level thereby actualizing the ideal of an informed patient. However, transparency and explainable systems are necessary (Bjerring and Busch 2021). Otherwise, the system becomes a threat towards the ideal of human-centered reproducible and understandable science as well as patient-centered care (Bjerring and Busch 2021).

Regulatory options

In current regulatory practice, the two principles of harmless-ness and effectiveness are used to assess the impact of novel interventions, e.g., in form of devices or medication. Regulations on medical devices (e.g., the EU Regulation 2017/745, or the German Medizinproduktegesetz) generally require producers to demonstrate the harmlessness of their products for patients or in case of expected harm (e.g., in radiation therapy) a risk mili-
gation and reduction strategy. In contrast, licensing for medication follows the framework of effectiveness in a series of medical studies determining a safe clinical dose (phase I), assessing side effects and efficacy (phase II), and ultimately demonstrating effectiveness (phase III) (Müllner 2005). Substances not demonstrating effectiveness can still be marketed, but under different legal frameworks as cosmetics or foods, not as medications.

With diverse pathways to choose, it is not surprising that AI companies pursue different legal strategies. Some AI tools for dermatological diagnosis already underwent the medical devices path of demonstrating harmlessness (e.g., A.S.S.I.S.T. (Online-Doctor 2022)). Others are marketing their products as providing simple non-medical services “not intended to perform diagnosis, but rather to provide users the ability to image, track, and better understand their moles” (AI Dermatologist 2023). These are strategies rather common in emerging markets. However, regulators should be aware of potential impacts applying the principles of harmlessness vs. effectiveness. When assessing AI performance, established parameters such as sensitivity, specificity, and precision should be complemented by a critical appraisal of biases and risks of the respective learning cycles and databases (Wehkamp et al. 2023). In this crucial moment regulators should align themselves with these developments and shape the legal landscape concerning safe and effective AI technology.

Besides legal regulation, medical guidelines (Leitlinien) systematically synthesize current knowledge based on clinical evidence. Balancing harmlessness and effectiveness, they give recommendations for action to medical practitioners without being legally binding. AI tools are currently not part of medical guidelines. However, given promising experimental results, guideline developers soon need to adopt a stance on AI tools. Here, critical assessment of the evidence is an important first step for including or excluding AI tools from recommendations. Successful RCT studies in image recognition should be validated in medical practice to arrive at a realistic assessment of performance and practical fit. With AI tools discussed in medical guidelines, clinicians will have more guidance to include or willfully exclude them in their practice. Recommending AI in medical guidelines would also entail clinical malpractice not to use the tools unless the patient agrees with below standard care (Thissen 2021). After all, responsibility for medical interventions lies with the human doctor and the informed patient as scenarios 1 and 3 show.

Ethical conundrums and agency constellations

Beneficence and non-maleficence, autonomy, fairness, and responsibility are among the guiding ethical principles discussed in healthcare provision (Beauchamp and Childress 2001). Complying with these principles is paramount for AI systems to integrate well into the healthcare system. For instance, enhancing autonomy means that patients should be given a choice to agree or disagree with the use of AI systems in their diagnosis procedure without negative consequences, i.e., higher health insurance premiums that would shift the burden of responsibility solely towards the patients and challenge the solidarity principle in health insurance (Böning et al. 2019). Fairness in this context would mean equal access to enhanced diagnostic procedures such as AI (WHO 2021). In the following, we will focus on responsibility, particularly regarding doctors and their decisions since diagnosing is primarily a task for medical professionals.

Legally and ethically, doctors are responsible for their medical decisions, and they are held accountable for malpractice and negligence. This strong belief in assuming responsibility stands in contrast with the opaque and thus fascinating nature of AI systems outlined above. Conflicts arise when an AI system provides a different interpretation compared to the dermatologist. With responsibility clearly being attributed to the human actor, the dermatologist is faced with the difficult task to reconcile their own beliefs with discordant input from the AI system. Figure 1 shows a contingency table with the dermatologist’s and the AI’s diagnosis expanding the ethical discussion by Tupasela and Di Nucci (2020) with a temporal dimension.

The concordant cases are straightforward. In dissenting cases, the serious diagnosis melanoma is likely dominant and guiding in the first instance, irrespective of the information source. The responsible dermatologist will likely escalate the diagnostic process ‘to be on the safe side’ and excise the suspicious skin lesion. This leads to a general increase in operations and a consequent influx in associated healthcare costs in a field where already only 1 in 10 operations identifies a case of disease (Petty et al. 2020). Given a learning curve with the AI system assumed to be even slightly more accurate than the dermatologist, the AI system’s appraisal over time becomes the dominant assessment irrespective of the diagnosis. In that case, the dermatologist’s

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>Human dermatologist</th>
<th>AI system</th>
</tr>
</thead>
</table>
| Melanoma         | Dissent
| Initially: melanoma
| Later: melanoma  | Concorde
| Harmless mole    | Initially: harmless mole
|                   | Concorde

Fig. 1: Contingency table with human dermatologist and AI system as actors diagnosing a mole as either melanoma or as harmless. In the top left and bottom right cell both actors agree. In the top right cell in a first instance the more dangerous diagnosis melanoma takes precedence. As the AI system proves more reliable than the dermatologist over time this diagnosis is upheld. In the bottom left cell, the more dangerous diagnosis melanoma also takes precedence. However, this is overwritten later by the AI system. Both dissenting cells create cognitive dissonance for the dermatologist. Source: authors’ own compilation
competence is depreciated, they are “confined to the role of a recording device” (Ellul 1964, p. 93), however, whilst assuming responsibility, liability, and accountability for the decision. Ethical conundrums arise challenging the agency of the dermatologists who do not fully understand the AI ‘decision-making’ process. Explainability and critical reflection of AI tools are necessary for the dermatologists to experience self-efficacy and interpret the results adequately (Bjerring and Busch 2021). This is particularly the case, as many profit-oriented companies might sense a chance in this emerging market by conveying the appearance of a functioning and accurate system.

Conclusion

We conclude that AI tools for diagnosing melanoma potentially provide convincing benefits for the healthcare system in terms of efficiency gains, more adequate resource allocation, health literacy and empowerment for patients, or more accurate diagnoses and better health outcomes. Promising experimental results must be validated in clinical practice. Three scenarios demonstrated applications of AI tools for diagnosing melanoma in different settings for different purposes.

However, the following necessary conditions must be fulfilled: AI tools must perform reliably with sufficient specificity and sensitivity; they must be transparent regarding the analytical processes and outcomes; and they must be accepted by patients, doctors, and other stakeholders. Moreover, AI tools need to adhere to ethical values of beneficence, non-maleficence, autonomy, fairness, and responsibility to protect dignity of all human actors involved. That includes AI tools being a support rather than replacement for human actors. Even considering all aspects above, cognitive dissonance in decision-making and competency shifts for dermatologists have to be expected particularly when the AI systems demonstrate superiority.

Based on these analyses we suggest technology assessment studies for AI in diagnostics to analyze the application contexts and consequences for the multiple stakeholders involved. Experimental studies focusing on performance should be complemented by observation studies in realistic settings of clinical practice. Regarding regulation, a nuanced debate about the underlying frameworks and an analysis of their consequences for accepting AI in diagnostics is needed.

Funding • This work received no external funding.
Competing interests • The authors declare no competing interests.

References

AI Dermatologist (2023): AI Dermatologist Skin Scanner. Available online at https://ai-derm.com/, last accessed on 03. 01. 2024.
Ayer, Jessica; Falloon, Caitlin; Ravi, Ayingaran; Ha, Phil; Le, Suong (2023): A model-based and prototype chatbot for increasing health literacy of patients with decompensated cirrhosis. Usability study. In: JMIR Human Factors 10 (1), p. e42506. https://doi.org/10.2196/42506
Petty, Amy et al. (2020): Meta-analysis of number needed to treat for diagnosis of melanoma by clinical setting. In: Journal of the American
Pham, Tri-Cong; Luong, Chi-Mai; Hoang, Van-Dung; Doucet, Antoine (2021): AI outperformed every dermatologist in dermoscopic melanoma diagnosis, using an optimized deep-CNN architecture with custom mini-batch logic and loss function. In: Scientific Reports 11 (1), p. 17485. https://doi.org/10.1038/s41598-021-96707-8
Saginala, Kalyan; Barsouk, Adam; Aluru, John; Rawla, Prashanth; Barsouk, Alexander (2021): Epidemiology of melanoma. In: Medical Sciences 9 (4), p. 63. https://doi.org/10.3390/medsci9040063
Schreier, Jan; Genghi, Angelo; Laaksonen, Hannu; Morgas, Tomasz; Haas, Benjamin (2020): Clinical evaluation of a full-image deep segmentation algorithm for the male pelvis on cone-beam CT and CT. In: Radiotherapy and Oncology 145, pp. 1–6. https://doi.org/10.1016/j.radonc.2019.11.021


**DR. JAN C. ZÖLLICK**
has been a researcher at the Institute of Medical Sociology and Rehabilitation Science at Charité – Universitätsmedizin Berlin since 2017. He conducts research and teaches about the opportunities and risks of technological innovations for health-care.

**PROF. DR. HANS DREXLER**
has been the Chair for Occupational, Social, and Environmental Medicine since 2000. He is a dermatologist and occupational physician. He focuses on the social-medical consequences of malignant skin tumors.

**DR. KONSTANTIN DREXLER**
has been working as a physician at the Clinic and Polyclinic for Dermatology at the University Hospital Regensburg since 2016. His professional and scientific focus is on the diagnosis and treatment of various skin tumors.

---

**GAIA Masters Student Paper Award**

The international journal GAIA – Ecological Perspectives for Science and Society invites Masters students to participate in the 2025 GAIA Masters Student Paper Award.

**Submission guidelines and more information:**
www.oekom.de/zeitschriften/gaia/student-paper-award

**Deadline for submission:** November 25, 2024.

The winner will be selected by an international jury and will be granted a prize money of EUR 1,500 endowed by the Selbach Umwelt Stiftung and Dialogik gGmbH, as well as a free one-year subscription to GAIA, including free online access. The winner may also be encouraged to submit his or her paper for publication in GAIA.