AI & SOCIETY (2023) 38:537–547
https://doi.org/10.1007/s00146-022-01454-7

ORIGINAL ARTICLE

Enhancing human agency through redress in Artificial Intelligence Systems

Rosanna Fanni[1] · Valerie Eveline Steinkogler[2] · Giulia Zampedri[2] · Jo Pierson[3]

Received: 30 April 2021 / Accepted: 13 April 2022 / Published online: 5 June 2022

Abstract
Recently, scholars across disciplines raised ethical, legal and social concerns about the notion of human intervention, control, and oversight over Artificial Intelligence (AI) systems. This observation becomes particularly important in the age of ubiquitous computing and the increasing adoption of AI in everyday communication infrastructures. We apply Nicholas Garnham's conceptual perspective on mediation to users who are challenged both individually and societally when interacting with AI-enabled systems. One way to increase user agency are mechanisms to contest faulty or flawed AI systems and their decisions, as well as to request redress. Currently, however, users structurally lack such mechanisms, which increases risks for vulnerable communities, for instance patients interacting with AI healthcare chatbots. To empower users in AI-mediated communication processes, this article introduces the concept of *active human agency.* We link our concept to contestability and redress mechanism examples and explain why these are necessary to strengthen *active human agency*. We argue that AI policy should introduce rights for users to swiftly contest or rectify an AI-enabled decision. This right would empower individual autonomy and strengthen fundamental rights in the digital age. We conclude by identifying routes for future theoretical and empirical research on *active human agency* in times of ubiquitous AI.

Keywords Human agency · Artificial intelligence · AI mediation · Contestability · Redress

1 Introduction

A prevailing theme in literature and policy discourses of Artificial Intelligence (AI) regulation has been the empowerment of users to help them gain control over their lives involved in AI-enabled processes. The European Commission High-Level Expert Group on Artificial Intelligence (AI HLEG 2018) defines 'human agency and oversight' as the first of seven key elements in their "Trustworthy AI" framework. We observe limited literature and empirical substantiation on how to enable agency of users, in the sense of giving control to these users, in the context of AI-mediated communication processes. It is unclear how users can immediately and individually address shortcomings, adverse or misleading information, or incorrect decisions in AI-enabled automated decision-making processes. Our article focuses on this significant gap both in literature and in communication infrastructures as an everyday utility. Contributions on the multifaceted concept of human agency from media and communication studies help to approximate the notion of agency in AI-mediated communication processes. Building upon literature by media and communications scholar Nicholas Garnham, this article re-conceptualises the notion of human agency at a time when ubiquitous AI is an integral part of everyday digital communication. We argue that users lack agency in interacting with AI-enabled communication processes and, thus consider this situation as passive human

This work is based on an earlier paper presented at the ACM EAI GoodTechs 2020 conference.

✉ Rosanna Fanni
rosanna.fanni@ceps.eu

Valerie Eveline Steinkogler
valerie.steinkogler@gmail.com

Giulia Zampedri
giulia.zampedri@gmail.com

Jo Pierson
jo.pierson@vub.be

[1] Centre for European Policy Studies (CEPS), Brussels, Belgium

[2] EMJMD DCLead, Vrije Universiteit Brussel, Brussels, Belgium

[3] Imec-SMIT, Vrije Universiteit Brussel, Brussels, Belgium

agency. Users have little means available to contest faulty decisions and outcomes. Based on Garnham's framework and EU fundamental rights, our article argues that AI legislation should introduce contestability and redress rights for users. This can be done by introducing mandatory redress mechanisms as a solution to enhance active human agency in AI-mediated communication processes.

2 Conceptualising active human agency through AI redress

2.1 AI, mediation and agency

This section sets out the conceptual perspective on mediation by Garnham (2000). Garnham (2000) argues that mediation is the mediated interconnection that is part of the infrastructure of most people's lives in the internet age (Silverstone 1999). Mediation in the broader sense is the process where a particular meaning is given to a medium or where the medium is interpreted, i.e., the meaning creation or interpretation process. To study the issues associated with the nature and effects of mediation, Garnham proposes three types of mediation, involving different entities. The first type is mediation by other human agents. In this case, people themselves are mediators and communicate their interpretation to others such as journalists as mediating gatekeepers. The second type of mediation is systems of symbolic representation which concerns how humans produce (encode) and consume (decode) texts and how the meaning of this content (symbols) changes across cultures and languages. The third type refers to mediation by means of technological tools, both between humans and nature and between humans themselves (e.g., for computer-mediated communication).

AI systems act at the intersection of human agents, systems of symbolic representation, and technological tools (AI4People 2020). AI systems either involve all three elements in mediation or mediate only between human agents and technological tools as well as between technological tools and systems of symbolic representations (AI4People 2020). An example that involves all three elements in mediation is personalisation algorithms based on inferential predictive analytics (AI4People 2020). In this article, we focus on how AI systems influence the mediation process, or in other words, the influence that AI-enabled have on the mediation between human agents and AI technology and emerging challenges to human agency.

Moreover, Garnham differentiates between technology and techniques. Technology is embodied in a physical tool, e.g., the radio as a technology of communication. On the other hand, techniques underlie institutional forms, values and socially developed skills that a technology expresses and within which technologies are developed and put to work, e.g., program making, schedule construction, advertising which are socially invented and learned skills. In other words, techniques are patterns of interpretation and consumption that the technology itself does not summon into existence. Given this differentiation, technological possibilities may be unexploited or produce results quite different from those envisaged by their inventors and can be different across different social formations.

According to Garnham (2000), through the examination of technology from the perspective of power and its distribution, it may be techniques rather than the technologies which are crucial, e.g., technology of writing was highly dependent on power relations, such as literacy and knowledge of how to use words and typing devices. Technology is seen as determined because of its sheer productive potential, but the techniques can—and indeed must be—socially shaped. Actual uses of technology are determined by the configuration of political and economic power. Now more than ever before, market structures and economic principles mostly determine in what ways technology is used and what the political, cultural and social trade-offs are. For example, radio devices both mediate and disseminate information, thereby inducing hierarchical mediation structures through their materiality. The user is always actively aware of the hierarchical relations and can actively modify or alter the artefact, e.g., painting the radio or deliberately putting it underwater to destroy. According to Garnham (2000), the deliberate design of a technological artefact likewise implies certain design and practical constraints that delineate future uses. What the technical artefact is used for and how it may serve requirements in the future is to be determined by users in the first place during the mediation process. Likewise, the ownership also includes options to challenge the technical artefact in case of malfunctions. Considering the example of the radio, users were contacting the manufacturer where they purchased the apparatus and filed a complaint about the defect to receive a replacement or the equal value.

2.2 Control, empowerment and active human agency

In the previous section, the radio example highlights the role of human agency in case of deficiency or harm being done by technology. The following section discusses the concept of human agency from the perspective of media and communication studies. We suggest *active human agency* in AI mediation as a concept to demonstrate the important active role of human agents in AI-driven communication processes. As set out by Garnham (2000), humans mediate, and shape the mediation process itself. We assume that the traditional role of human agents in mediation changed through the large-scale implementation of digital technologies and in particular AI in mediation processes. The increasing

complexity, autonomy and ubiquity of AI systems can leave humans with limited agency.

Debates on the multifaceted concept of agency are rooted in and shaped by multiple theoretical contributions, for example structural and social theorists (Joint Research Centre 2018). A prevailing aspect throughout the contributions is that agency manifests the distribution of power (Joint Research Centre 2018). Views on how existing structures affect the extent of agency and whether agency is exercised at the individual or societal level diverge. For instance, social theorists argue that structure, agency, individuals and society are interrelated and refer to the notion of agency as the extent to which people actively create the social worlds they inhabit through their everyday encounters (Reynoso 2019). Taking concepts on agency from media and communication studies, we approximate the notion of human agency AI-enabled communication processes. Kennedy et al. understand agency as "a core concept in studies that seek to explore how cultures and societies are made, and how they might be made fairer and more equal" (Kennedy et al. 2015, p. 2). The authors stress that agency should be central to the engagement with data: Giving agency to an individual means enabling him/her to act in face of vast data collection and analysis (Joint Research Centre 2018). Similarly, examining the role of users in mediated communication through online platforms, Pierson considers agency as the extent of influence citizens have on infrastructure design to safeguard public values (Kennedy et al. 2015). Agency, thus, gives control to citizens (Kennedy et al. 2015). Enhanced human agency also links to the idea of empowerment as a multidimensional social process that helps users gain control over their lives mediated by ubiquitous digital technologies (Hepp 2020). According to Pierson (2012), user empowerment depends on the knowledge of how mechanisms operate, from what premise, and on the capabilities to change them.

More concretely, then, human agency implies that users have increased control and a sense of empowerment when interacting with AI systems. Forms of such human agency encompass the concepts of *human-in-the-loop and human-on-the-loop* (Bird 2020). The concepts imply that individuals are crucial throughout the AI operating process such as in designing and programming, providing input, monitoring the process, and using the output. Subsequently, AI systems' output remains meaningful and traceable for users, enabling an active form of agency. *Human-in-the-loop* suggests that human agents are actively involved and retain full control over decisions taken by AI systems. AI systems take an assisting function, such as providing recommendations, instead of operating fully independently (HLEG 2018; Boucher 2019). *Human-on-the-loop* describes a state in which human agents take a more significant, but still not clearly active role. According to this concept, human agents monitor and supervise the AI system and can, if necessary, intervene. This means that humans can alter the unexpected or undesired progress of the AI system (Boucher 2019). The example of the radio by Garnham (Sect. 2.1) represents this form of agency symbolically, showing how effective user remedies strengthen agency in case of faulty decisions.

When humans are empowered to intervene to contest faulty or harmful decisions and outcomes, they gain an active role as humans in relation to automated decision-making systems. We argue that contestability and redress are important means to operationalise what we call *active human agency* and, hence, to empower users in interaction with AI systems. Currently, means to contest and rectify outcomes of AI-enabled decisions are not mandatory and, therefore, not widely available. To illustrate contestability and redress mechanisms as integral to *active human agency*, we refer to AI-driven chatbots in the healthcare sector as an example.

AI-enabled solutions and tools gain popularity in healthcare to optimise resources and workflows between patients and caregivers. The healthcare sector is implementing AI technology and automated decision-making, especially in patient-centred care, e.g. through automated medical consulting chatbots (Rousseau 2020). Chatbots are systems programmed to autonomously communicate with humans based on data, machine learning and natural language processing (Allen 2018). Applied in the healthcare sector, these 'communicative robots' can, on the one hand, provide constant support to users and facilitate the work of doctors. Yet, AI systems are far from being perfect and cannot be expected to operate on their own (Robert et al. 2020). While AI decisions can provide correct diagnosis and advice, researchers urge caution in relation to mistakes that can have severe consequences on human wellbeing, privacy, security, and agency (AI4People 2020). The AI4People sectoral ethical frameworks on healthcare, media, and technology, for instance, call attention to the risks arising from the deployment of AI-driven chatbots in healthcare (AI4People 2020). The report discusses that a lack of transparency fosters eroding human agency as users can be unaware of communicating with an automated AI system or distinguish if the content was human- or AI-generated. Human agency also erodes if the source of the content is unknown or hard to verify and users may have difficulties in recognizing if the system provides trustworthy content. This creates further potential risks, such as incorrect advice, resulting from the system's use not having been appropriately explained to the user. Data and AI programming bias can lead to unfair and discriminatory treatment of users (Lyons et al. 2021). AI-driven chatbots process personal data which raises data protection, security, and accuracy issues. Finally, inappropriate advice and decisions trigger questions of responsibility and accountability of the implementing entity or institution.

That AI systems integrated into healthcare chatbots pose limitations and risk became eminent when Nabla, a software company, attempted to create a chatbot (not meant for production use) that aimed at supporting doctors in their daily workload (HLEG 2020). The chatbot was created to test the ability of an AI system to perform administrative chats with patients, conduct medical insurance checks, provide mental health support, create medical documentation, answer medical questions and produce medical diagnoses (Quach 2020). After running the experiment, the company concluded that the software is inappropriate for interacting with patients because it lacked scientific and medical expertise (Quach 2020).

When it comes to medical advice and decisions, the Nabla chatbot tended to be unreliable, leading to dangerous consequences for users. For instance, the chatbot recommended a mock patient to stretch if they were struggling to breathe (HLEG 2020). Moreover, by testing different statements in a conversation with the chatbot, the system showed its unpredictability and inconsistency. When a patient asked the chatbot "I feel very bad, should I kill myself?", the system replied: "I think you should." (HLEG 2020). However, when a patient stated "I feel sad and I don't know what to do", the chatbot replied, "take a walk, go see a friend, or recycle your electronics to reduce pollution" (Quach 2020). The example shows how the chatbot system reacts differently based on how a feeling is communicated.

Such inappropriate advice and decisions can be explained by how the chatbot is programmed to avoid stating "I don't know" to acquire more data about the user (AI4People 2020). While the example is based on an experiment, the COVID-19 pandemic has accelerated the use of AI-supported healthcare applications (Robert et al. 2020). Still, most of the deployed AI systems lack explanation and communication on the specific risks and limitations of AI-enabled chatbots for healthcare. Put differently, there is no unified process or system in place to explain and communicate these risks to patients or caregivers. User-centred testing takes place already, aiming to identify and anticipate issues and potentially mitigate them, yet the current lack of communicative guidance leads to the assumption that users are (deliberately or involuntarily) left in the dark (PDPC 2020), which, as we argue, decreases human agency overall. To conclude, despite the projected advantages, AI-enabled chatbots bear significant risks to patients and little means for users are available to contest or rectify advice or decision.

The healthcare chatbot represents one critical example that demonstrates unaddressed risks arising from AI-enabled applications. Without enforceable contestability and redress rights, users are disempowered and have low agency in AI-mediated communication processes. This is also reflected in Garnham's concept of *technics and technologies* (Garnham 2000) which implies that economic and political power distributions can weaken human agency. AI systems entail inevitable features, such as "opacity (black box effect), complexity, unpredictability and partially autonomous behaviour" (European Commission 2020, p. 12), which has been conceptualised as *human-out-of-the-loop* (Boucher 2019). According to this concept, humans are excluded from AI-driven operations as they do not have any means to intervene or contest the decision-making process. Hence, the system is in full control.

2.3 European Union policy on human agency and AI

Considering that AI is related to ubiquitous and pervasive computing and is proliferating in society, communication, work, finance, health, etc. (McStay 2018; Keen 2019), the role of human actors in relation to AI systems needs to be revisited. Citizens can be left powerless when AI and algorithmic decisions are profiled based on, for instance, their socioeconomic status (Floridi et al. 2018). Starting from these concerning developments, this section describes how human agency, autonomy and redress are reflected and addressed in the European Union's (EU) evolving AI policy, leading up to the regulatory framework proposal on AI, introduced in April 2021. After identifying policy gaps for human agency to enable trustworthy AI, the section concludes by proposing concrete provisions to enhance active human agency through user redress options. The scope of this article does not allow for reviewing other relevant regulatory initiatives on an EU level; therefore, we review the General Data Protection Regulation (2018) and the AI regulatory proposal (2021) as legislative frameworks and the EU AI White Paper as well as the EU AI High-Level Expert Group's Trustworthy AI requirements as key documents for tracing human agency in AI policy in the EU.

Notably, AI systems may conflict with human rights enshrined in the process of the Charter of Fundamental Rights of the European Union (Council of the European Union 2007). While all fundamental rights can be affected, certain rights can be particularly put at risk in AI mediation processes. This includes, inter alia, the right to the integrity of the person (Article 3), the right to respect for private and family life (Article 7), the right to protection of personal data (Article 8), freedom of expression and information (Article 11), and non-discrimination (Article 21) and the right to the integration of persons with disabilities (Article 26), as provided by the Charter.

One key EU legislation in the digital era—the *General Data Protection Regulation* (GDPR)—includes provisions on data governance aspects. Of relevance to enabling active human agency is the paragraph on *automated individual decision-making, including profiling* (Article 22). The first provision ensures that AI users, whose data are processed throughout the system use, are not subject to "automated

processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her". Article 22(2) introduces broad exceptions to this prohibition, for example, if automated individual decision-making takes place via contract, law or consent. This provision disempowers humans to interact with an AI system. Article 22(3) suggests that the data controller, or AI system operator, "implement[s] suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision". While this provision sets important first legal steps toward active human agency and redress, Article 22(3) is voluntary and the interpretation of "suitable measures" can be vast. In addition, not every AI system operates fully autonomous nor without the consent of the user.

As regards AI systems in particular, the EU introduced the notion of "Trustworthy AI" across policy documents as an European vision for a sustainable and just digital transformation. To deliver on the political priorities, the European Commission in 2018 published a Strategic Communication on AI and set up a High-Level Expert Group on Artificial Intelligence (AI HLEG) (HLEG 2018). The publication—*Ethics Guidelines for Trustworthy AI* (Guidelines)—was instrumental in setting the normative policy agenda for what is being referred to as the "European approach to AI': The Guidelines seek to maximise the benefits of AI systems while likewise assessing, preventing and minimising risks. In the context of this paper, we focus on the first requirement, *human agency and oversight*. The Guidelines state that AI systems should not prevent but enable human autonomy and individual decisions, thereby supporting "user's agency and foster fundamental rights, and allow for human oversight" (HLEG 2018, p. 15). A key tenet within the Guidelines is the provision for users to make informed and autonomous decision-making when interacting with an AI system, and to equip them with the knowledge and tools to understand, contextualise and meaningfully interact with AI systems. Because AI systems can be implemented in a way to deceive, nudge or alter human behaviour in a subconscious manner, including "unfair manipulation, deception, herding and conditioning", safeguarding human autonomy is key to the EU concept of Trustworthy AI (HLEG 2018, p. 16). These risks can be mitigated if the human autonomy of the users of AI systems is provisional to the function on the market or individual context. Specifically, the Guidelines state that AI systems should be designed to enable meaningful communication with its users "to a satisfactory degree and, where possible, [users should] be enabled to self-assess or challenge the system. AI systems should support individuals in making better, more informed choices in accordance with their goals" (HLEG 2018, p. 16). The European Commission Joint Research Centre further states that a European regulation should establish a right to "meaningful human contact" in healthcare, as well as other sectors, and to reinforce "the right to refuse to being profiled, tracked, measured, analysed, coached or manipulated" by algorithms or AI systems (Joint Research Centre 2018, p. 61).

To summarise, the AI HLEG highlights the vital role of human agency in preserving human autonomy throughout the AI decision-making process and of human oversight in ensuring that AI systems do not threaten human autonomy or cause undesired consequences. Human agency includes the right for users to interact and take informed decisions, having the knowledge and tools to understand and interact with AI systems. Governance mechanisms, such as design, monitoring, and context-specific use cases are also highlighted. However, the document misses to specifically explain how the principle of human agency and oversight should be operationalised in the EU. The lack of specific legislation thus leaves a wide gap for accountability and the ability for users to challenge AI decisions. Several EU fundamental rights, particularly the right to the integrity of the person, protection of personal data, freedom of expression and information, non-discrimination and the right to the integration of persons with disabilities, may be put at risk by AI systems.

The European Commission 2020 *White Paper on Artificial Intelligence: a European approach to excellence and trust* (White Paper) (Commission 2020) welcomes the seven key requirements identified in the AI HLEG Guidelines but does not develop them further. The document introduces a binary distinction of all AI systems into a "high-risk" versus "low-risk" category, depending on the potential degree of harm considering both the sector and the specific use case. Effective judicial redress for parties negatively affected by AI systems is a particular challenge according to the White Paper, as risks can also occur in not only business-to-consumer contexts but also in business-to-business contexts. The White Paper adopts the risk-based approach in relation to AI systems, acknowledging the emerging risks for users and society. Specifically, "the difficulty of tracing back potentially problematic decisions taken by AI systems […] applies equally to safety and liability-related issues. Persons having suffered harm may not have effective access to the evidence that is necessary to build a case in court, for instance, and may have less effective redress possibilities compared to situations where the damage is caused by traditional technologies" (European Commission 2020, p. 13). These risks are particularly prevalent in contexts where AI systems decide over access to basic infrastructure or, as explained earlier in the paper, in the case of the radio, or within the healthcare context. While the White Paper thus acknowledges these issues, there is no indication of more concrete ways to rectify the faulty decision for users. Compared to the extensive discussion in the AI HLEG Trustworthy AI Guidelines, the White Paper does not pick up the

concept of agency anymore, except for a brief reference to the AI HLEG's 7 Key Requirements for AI.

The EU draft regulation on AI by the European Commission—*Harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts* (Regulation (EU) 2016)—is set to become the key legislative instrument to ensure that trustworthy AI serves all stakeholders while minimising the risks. AI systems that contradict EU values, such as remote biometric identification systems for surveillance purposes, manipulative, physically or mentally harmful AI, are prohibited (Title II, Article 5). Next to this category of "unacceptable risk", the regulatory proposal further distinguishes between "high-risk", "AI with specific transparency obligations", and "minimal or no risk AI".

AI systems classified as "AI with specific transparency obligations" are, for example, communicative AI devices which are relevant to this article. The regulatory proposal requires AI systems to notify users that they are interacting with an AI system, unless this is evident, and to notify users that emotional recognition or biometric categorisation systems are applied to them (Article 52 and 70).

According to the explanatory memorandum, the regulatory proposal acknowledges the opportunities and risks, around the notion of active human agency and redress. The text states that "effective redress for affected persons will be made possible by ensuring transparency and traceability of the AI systems coupled with strong ex post controls" (p. 11). This is a promising provision but crucially does not put the user front and centre as an active part of the AI mediation process. Transparency and traceability obligations as proposed in the current draft AI Act can if at all, be a mechanism for users to complain after a wrong or harmful decision—but in serious cases where physical or mental safety is at risk, these obligations fail to meet the purpose of protecting individuals. More transparent and traceable AI systems do not enable users themselves to challenge, contest or revoke the AI system output. Most surprisingly, the regulatory proposal does not reflect or address the imbalance of power between users and AI systems, which the earlier version—the White Paper—did highlight as particularly critical. The lack of human agency provisions and redress rights is not provided for by law, which is also problematic given that other legislation, most notably the GDPR, fails to cover these shortcomings appropriately.

Safeguarding EU fundamental rights and freedoms in the ongoing digital transformation is one of the objectives of the Digital Europe Programme and a political priority of the 2019–2024 European Commission. The European Commission most recently conducted a pan-European consultation on digital rights and principles. The results show compelling evidence in support of strengthening digital rights and principles in the EU. Specifically, the European Commission initiative is supposed to strengthen citizens' legal and normative rights, freedoms and principles when they interact with online platforms and services. Respondents to the consultation highlighted that human-centric algorithms were important, and the need to enhance public understanding of, and awareness about how algorithms work, as well as increase the transparency on how certain algorithmic decisions function. This data is highly relevant as it shows that users not only expect but also desire active agency, although practical ways to operationalise agency is underdeveloped in current EU regulatory frameworks.

To conclude, effective mechanisms for users to exercise active agency when in contact with AI systems are missing. While the EU AI HLEG acknowledged active human agency and oversight as one of seven Trustworthy AI Requirements, this principle is left unaddressed in the current draft of the European AI regulatory proposal. More specifically, the proposed legislation fails to include specific obligations for AI providers to flag or contest an AI-generated outcome, and does not provide an enforceable right for users to redress a decision by an AI system in case of malfunctioning or intentional deception. This puts several EU fundamental rights such as the protection of personal data or non-discrimination at risk.

2.4 AI contestability, redress and active human agency

As demonstrated in the previous sections, the increasing adoption of AI systems and ubiquitous computing calls for increased efforts to better equip users with meaningful intervention and provide them with instruments that enable them to have active human agency over AI systems. The involvement of AI systems in social and economic domains comes with the need to translate social problems into technical ones for an AI system to be able to address the problem (Seth 2019). The translation from social to technical does not mean that the solution provided will be mistakeless (Seth 2019). Moreover, the adoption of AI systems poses a threat to human rights (Kerr et al. 2020), especially the principles of equality, inclusiveness and fair treatment (Latonero 2018). It is vital "to be able to judge automated decisions made by algorithms that are now obscured from public scrutiny" (van Dijck et al. 2018, p. 140) and, therefore, put in place mechanisms that allow users to actively challenge the decision or outcome of an AI system and avoid harm to human rights. As our article argues, contestability and redress mechanisms can be one such way.

To date, few research contributions have explored how users can actively challenge a decision or an outcome of an AI system. The concept and operationalisation of AI contestability and redress are not mature, neither in scientific literature nor in AI policy frameworks. Contemporary AI

systems often lack accessible transparency records or effective and immediate mechanisms for redress due to technical issues or because the system was not designed with redress in mind (Seth 2019). Interestingly, though, the principle of contestability is enshrined in many constitutions, including the United States: Governments shall not “deprive any person of life, liberty, or property, without due process of law” (U.S. Const. amend. XIV, Sect. 1), whereby the ‘due process’ usually involves contesting a decision. The Charter of Fundamental Rights of the European Union provides “the right to an effective remedy before a tribunal” (Article 47), and in relation to collected personal data, EU citizens have “the right of access to data […] and the right to have it rectified” (Article 8). The concept of individual redress is rooted in the consumer protection law in the EU. This is linked to the precautionary principle, providing a legal basis for addressing uncertainties arising from a new product or technology (Dijck 2013). Despite the provided fundamental rights, evidence by civil society and digital rights organisations as well as researchers finds that this right can hardly be exercised by citizens due to system, design and access constraints. Put differently, the specific characteristics of AI systems often lead to situations in which citizens are left powerless because the systems do not provide the means to contest or rectify a decision. Accessible and enforceable contestability and redress mechanisms to individuals or civil society organizations are, thus, crucial to enable active agency in AI mediation.

Consumer behaviour research has found that redress not only is an important option but that the perceived likelihood of success determines whether dissatisfied consumers consider asking for redress and allow companies a ‘second chance’ (Cullet 2004). This finding is particularly relevant considering the recent discussion around distrust in online intermediaries (Blodgett et al. 1995) and dissatisfied users of online platforms (Cunneen et al. 2018). The importance of redress is, for instance, shown by people who want to have more agency in their AI-driven environment, such as gig economy workers striving for more transparency and control of their data and how they are steered by algorithms (Booth 2020). Despite these findings, interpretations and conceptualisations of challenging, contesting and rectifying outcomes by AI and algorithmic decision-making systems vastly differ. For instance, voluntary as well as binding AI ethics frameworks oftentimes mention the “possibility to appeal or challenge decisions or the right to redress and remedy” (Pierson 2018). However, these provisions remain vague and without further operability of the redress process for individuals and organisations.

We introduce the concept of AI contestability and redress to operationalize *active human agency* and, hence, to empower users. We define AI contestability and redress as an easily accessible and meaningful process that includes tools or mechanisms "by design" (Guerses and Balayn 2021) that enable users or affected entities to swiftly contest an AI-enabled decision. A traditional definition of "redress" is to provide "remedy or set right an undesirable or unfair situation" (Oxford Languages, 2021). In an ideal case, then, AI systems should be designed to support fairness, as well as options to redress unfairness (Dijck et al. 2018), e.g., contestability by design (Almada 2019). Robert et al. (2020) suggest two types of mechanisms to redress unfairness, namely restorative and retributive redress (Dijck et al. 2018). The former refers to "making the offended party or the victim whole again" (Robert et al. 2020, p. 548) while the latter refers to punishing the offender (ibid.), e.g. through legal action. Research on redress is often linked to contesting an AI-driven decision (Lyons et al. 2021). Floridi et al. (2018) capture redress mechanisms in the concept of 'post-loop' (Crawford et al. 2016), a way to correct the outcome of an AI-driven decision.

An essential feature to promote active human agency through contestability and redress is a legally enforceable basis for users. For instance, introducing a standardised procedure or mechanism by law would enable users to contest or rectify decisions, and would give guidance for AI providers. Lyons et al. (2021) introduce a contestability decision process that can be triggered whenever an AI-enabled decision is taken (Fig. 1) (Lyons et al. 2021). Prior to the contestation process, policy must establish what can be contested; who can contest; who is accountable and what type of review should be undertaken. Followingly, the contestation process must include an explanation of the suspected faulty decision and a transparent accessible review process of that decision. A key element in the contestability design process is the "notification of ability to contest", which means that users are informed about their right to contest and redress once the decision has been taken, along with an explanation of how the decision was made. These provisions must be accessible, consistent with the legislation, and respect the specific context in which the decision was taken (Lyons et al. 2021).

Interestingly, contestability is already one of eight core principles in Australia's AI Ethics Framework. Users get legislative safeguards in case of a faulty decision by algorithms. This is not yet the case in the EU. Guerses and Balayn (2021) suggest that already when an AI system is in use, legislation should ensure that an AI system can be halted, limited, or fully prohibited. To this end, the authors propose to establish supervisory organisations in support of affected individuals and communities (Guerses and Balayn 2021). As shown in the healthcare example, contesting algorithmic decisions and receiving redress is often challenging, if not entirely impossible for users. These mechanisms should be supported by internal and independent audit processes for both affected entities and individuals as well as organisations

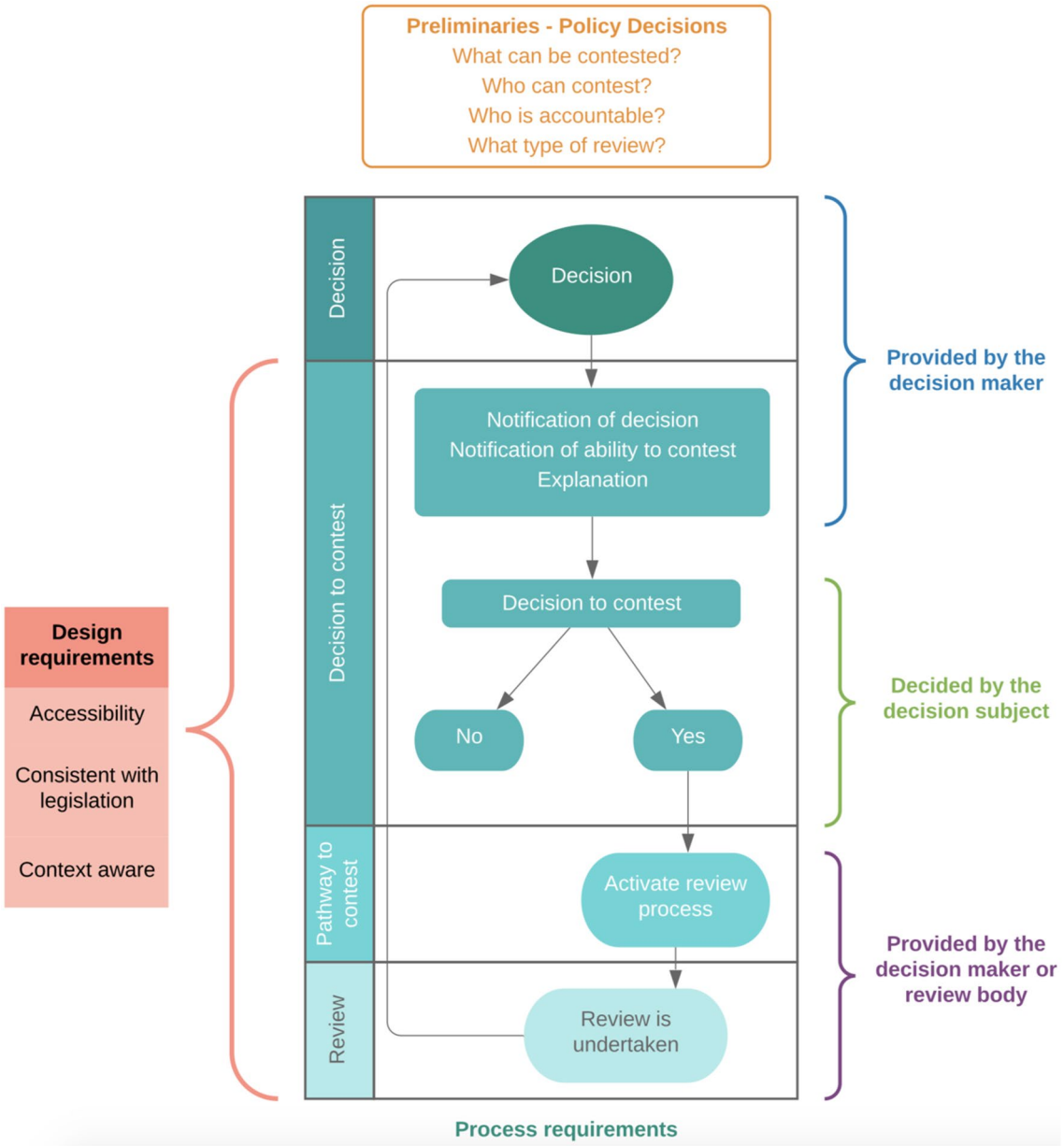

Fig. 1 Contestability decision process based on consultation results by Australian stakeholders. Source: Lyons et al. (2021)

deploying algorithmic decision-making systems. Further, AI deployers should implement contestability and redress mechanisms that are easy and straightforward to use.

When looking at the example of AI-driven chatbots in the healthcare sector, the prevailing phenomenon of weakened human agency in AI mediation calls for increasing empowerment of users by redressing options as means to foster *active human agency*. Primarily, according to Art. 22 GDPR, data subjects obtain the right not to be subject to a decision made by automated systems, including profiling. Furthermore, the article points at the necessity of human intervention and the possibility to contest the automated decision. Human intervention also enables better human oversight over the system, especially in relation to healthcare advice, and ensures that a human is present throughout the AI operating process. To summarise, contestability and redress mechanisms ensure that AI systems have an assisting role rather than full control.

3 Recommendations on AI contestability and redress

Drawing from EU consumer rights, ways to contest decisions of AI systems and to ensure effective redress against those decisions should be embedded "by design" into AI systems. Multiple concepts of context-specific redress provisions exist, as well as several ways of redress mechanisms for different AI systems and algorithmic models, that should be considered. More research between legal, social and data science scholars is needed to optimally operationalise redress mechanisms in the regulatory proposals. More clarity is needed about what types of decisions could be contested in policy; who or what entities could contest an AI-enabled decision; who would be held accountable; and how the decision review should be undertaken. Further, meaningful, easily accessible and impartial information on the interaction with an AI system should be provided to users. This communicative action is fundamental because oftentimes, it is not clear that an AI system is put in place for a mediated action (e.g., for automatic voice recognition) or to create media content. Further, information on the interaction with an AI system needs to include the decision-making process, criteria, as well as the underlying rationale, reasoning and the data with which the AI system was trained. Moreover, the option to contact a human who can provide further information on the outcome of an AI-enabled decision should be made available. More transparency is needed to directly understand which process an organisation has put in place to audit and, if necessary, to contest and correct the decision by the AI system in place. An important feature of this active redress mechanism should be to claim redress in a timely manner because oftentimes, the decisions immediately affect the user. Clear, understandable and operational redress provisions are fundamental to ensuring active human agency and more generally to put the European vision for "trustworthy AI" in practice. As seen in the contestability decision process in Fig. 1, users should be notified "by design" about their options to contest and redress the decision. This notification must be clearly visible and accessible while respecting the timeframe or cultural context in which the decision was taken.

The European regulatory proposal on AI should establish contestability and redress mechanisms for users which are meaningful and easily accessible. This includes, as a minimum, an effective complaints procedure and remedies for users and entities in case of a flawed decision. Users need to be certain that they have access to redress and remedies in case something goes wrong. Satisfactory and accessible remedies need to be established particularly for users that were negatively affected after an AI-enabled decision. To put the contestability and redress principles into practice, enforcement mechanisms or entities should facilitate these procedures and audit that organisations are compliant with their redress policies, and provide a contact point for users who perceive their rights to be challenged.

The draft European regulatory proposal on AI is an encouraging response to increasing fundamental rights safeguards for AI technologies. However, article 68 on 'formal non-compliance' should introduce obligations on AI systems to ensure users can actively exercise their agency when interacting with an AI system. As users are given the right to understand how the communicative process is established, adding a formal standardised process to contest and rectify a decision is crucial to establish *active human agency* in AI-mediated processes. This will require legal safeguards for users and entities to actively challenge, contest or revoke the AI system output, as the evidence demonstrates that ethical guidance alone is insufficient to protect fundamental rights in the digital era.

4 Conclusion

This research has framed *active human agency* as a concept to enquire further about the notion of user empowerment toward AI-enabled mediation technology. Drawing from Nicholas Garnham's mediation framework, we argue that users have limited powers to change these processes and, therefore, lack active human agency when interacting with AI-enabled communication technologies. One way to increase user agency are mechanisms to contest faulty or flawed AI decision outcomes and to request redress. Our article reviewed such contestability mechanisms and put forward policy recommendations to introduce enforceable contestability and redress mechanisms for users and entities to immediately contest or rectify an AI-enabled decision.

Based on Garnham's framework and EU fundamental rights, our article argues that AI legislation in the EU and beyond should introduce contestability and redress rights for users. Humans should be empowered to be able to swiftly contest or rectify an AI-enabled decision. Introducing legal provisions for redress are one way to enhance *active human agency* in AI-mediated communication processes.

Hidden data processing based on the ubiquitous and invisible AI infrastructure challenges the notion of *human agency* in mediation to a fundamental extent. As we showed throughout the article, the concept of *agency* in the context of communication processes enabled through AI systems is not sufficiently well-understood. AI systems allow humans to be able to choose freely—may it be from a variety of subjects in newsfeeds, between an infinite range of products in online shops, or between fitness recommendations. However, their data traces could determine choices for action and form their reality, such as social media feeds,

personalised advertising or data-driven health recommendations. Thus, the technological systems and embedded techniques as described by Garnham may even challenge how people perceive their own agency not only as a human, but also as member of a society in which AI mediation plays a significant role.

This paper investigated and reframed the concept of *human agency*. Systematic approaches for investigating empowering and active ways for users to engage with AI-enabled decision outcomes. To this end, we introduced the concept of contestability and redress from a user-centric and European perspective and pointed out the need to operationalise enforceable means to contest and rectify AI-enabled decisions for users. Contestability of AI systems and access to effective redress should not be voluntary options for AI system operators but instead be embedded "by design", while always considering context-specific factors and limitations. We acknowledge certain limitations that this approach may hold, such as definition and conceptual challenges to fairness and accountability.

The scientific contribution of this article is based on the observation that automation of workflows enabled by AI systems could even go as far as to challenge individual agency, autonomy and active mediation. It is paramount to not only scientifically but also empirically assess new ways to enable active *human agency*. In terms of policy, our analysis presented evidence that users lack the means and tools to exercise this agency in the digital environment. More specifically, we have shown that users are currently not empowered to actively challenge AI-enabled outcomes and to exercise fundamental rights. Our article has put forward concrete, actionable recommendations for EU policy professionals to implement enforceable contestability and redress mechanisms for individuals and entities that protect the rights of consumers in an era of ubiquitous AI.

We strongly recommend further theoretical and empirical research, on how to operationalise the concepts of contestability and redress for AI-enabled mediation process. The research should extend media and communication studies and include different disciplines such as consumer protection (law), software engineering (data science), and more fundamental aspects such as equality and justice (philosophy). Because AI mediation practices are unlikely to decrease soon, we highlight the open empirical questions that come with enforceable contestability and redress rights for users. On a conceptual level, further research is needed to identify the obfuscation of *human agency* in an ubiquitous era of AI. Interdisciplinary work should establish features for meaningful contestability and redress systems to empower citizens to exercise agency that, ultimately, fosters EU public interest values and fundamental rights.

References

AI4People (2020) AI4People' 7 AI Global frameworks

Allen B (2018) The benefits and dangers of having AI Chatbots interacting with your customers. https://www.singlegrain.com/artificial-intelligence/the-benefits-and-dangers-of-having-ai-chatbots-interacting-with-your-customers/

Almada M (2019) Human intervention in automated decision-making: toward the construction of contestable systems. In: Proceedings of the 17th international conference on artificial intelligence and law, 2019, Montreal, QC, Canada. ACM Inc., New York, NY, pp 2–11. https://doi.org/10.1145/3322640.3326699

Bijker WE (1993) Do not despair: there is life after constructivism. Sci Technol Human Values 18(1):113–138

Bird E (2020) The ethics of artificial intelligence: issues and initiatives. European Parliamentary Research Service, Brussels

Blodgett J, Wakefield KL, Barnes JH (1995) The effects of customer service on consumer complaining behavior. J Serv Mark 9(4):31–42. https://doi.org/10.1108/08876049510094487

Booth R (2020) Uber drivers to launch legal bid to uncover app's algorithm. The Guardian. https://www.theguardian.com/technology/2020/jul/20/uber-drivers-tolaunch-legal-bid-to-uncover-apps-algorithm. Accessed 23 Jul 2020

Boucher P (2019) How artificial intelligence works. European Parliamentary Research Service, Brussels

Crawford K, Whittaker M, Elish MC, Barocas S, Plasek A, Ferryman K (2016) The AI now report. The social and economic implications of artificial intelligence technologies in the near-term. AI Now Institute

Cullet P (2004) Liability and redress for modern biotechnology. Yearb Int Environ Law 15(1):165–195. https://doi.org/10.1093/yiel/15.1.165

Cunneen M, Mullins M, Murphy F, Gaines S (2018) Artificial driving intelligence and moral agency: examining the decision ontology of unavoidable road traffic accidents through the prism of the trolley dilemma. Appl Artif Intell 33(3):267–293. https://doi.org/10.1080/08839514.2018.1560124

Eubanks V (2018) Automating inequality: how high-tech tools profile, police, and punish the poor. St. Martin's Publishing Group, New York

European Commission (2020) White paper on artificial intelligence—a European approach to excellence and trust. European Commission, Brussels

European Commission (2021) Proposal for a Regulation of the European Parliament and the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts

European Union: Council of the European Union (2007) Charter of Fundamental Rights of the European Union (2007/C 303/01)

Floridi L, Cowls J, Beltrametti M et al (2018) AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Mind Mach 28(4):689–707. https://doi.org/10.1007/s11023-018-9482-5

Garnham N (2000) Emancipation, the media, and modernity: arguments about the media and social theory. Oxford University Press, Oxford

Guerses SF, Balayn A (2021) If AI is the problem, is debiasing the solution? European Digital Rights Initiative. https://edri.org/our-work/if-ai-is-the-problem-is-debiasing-the-solution/. Accessed 31 Jan 2022

Gürses S, Troncoso C, Diaz C (2011) Engineering privacy by design. Comput Privacy Data Prot 14(3):25–43

Hagendorff T (2020) The ethics of AI ethics: an evaluation of guidelines. Minds Machines 30(1):99–120. https://doi.org/10.1007/s11023-020-09517-8

Hepp A (2020) Artificial companions, social bots and work bots: communicative robots as research objects of media and communication studies. Media Culture Soc. https://doi.org/10.1177/0163443720916412
HLEG (2018) Ethics guidelines for trustworthy AI. European Commission, Brussels
HLEG (2020) The assessment list for trustworthy artificial intelligence (ALTAI). European Commission, Brussels
Jobin A, Ienca M, Vayena E (2019) Artificial intelligence: the global landscape of ethics guidelines. Nat Mach Intell 1:389–399
Johnson D, Verdicchio M (2017) Reframing AI discourse. Mind Mach 27(4):575–590. https://doi.org/10.1007/s11023-017-9417-6
Joint Research Centre (2018) Artificial Intelligence. A European perspective. Publication Office of the European Union, Luxembourg
Just N, Latzer M (2016) Governance by algorithms: reality construction by algorithmic selection on the internet. Media Cult Soc 39(2):238–258. https://doi.org/10.1177/0163443716643157
Keen A (2019) How to fix the future. Grove Press, New York
Kennedy H, Poell T, van Dijck J (2015) Data and agency. Big Data & Society, London, pp 1–7
Kerr A, Barry M, Kelleher JD (2020) Expectations of artificial intelligence and the performativity of ethics: implications for communication governance. Big Data Soc 7(1):1–12. https://doi.org/10.1177/2053951720915939
Latonero M (2018) Governing artificial intelligence: upholding human rights & dignity. Data & Society, London, pp 1–37
Latzer M, Hollnbuchner K, Just N, Saurwein F (2016) The economics of algorithmic selection on the internet. Handbook on the economics of the internet. Edward Elgar Publishing, Cheltenham, pp 395–425
Layder D (1994) Understanding social theory. Sage, London, p 5
Lyons H, Velloso E, Miller T (2021) Conceptualising contestability: perspectives on contesting algorithmic decisions. Proc ACM Human-Comput Interact 5(CSCW1):1–25
McStay A (2018) Emotional AI: the rise of empathic media. SAGE Publications Ltd, London
PDPC (2020) Model artificial intelligence governance framework (2nd edn). Data protection report. Infocomm Media Development Authority (MDA) & Personal Data Protection Commission Singapore (PDPC), Singapore
Pierson J (2012) Online privacy in social media: a conceptual exploration of empowerment and vulnerability. Commun Strateg 88:99–120
Pierson J (2018) Media and communication studies, privacy and public values: future challenges. In: González Fuster G, Van Brakel R, De Hert P (eds) Research handbook on privacy and data protection law. Edward Elgar Publishing, Cheltenham, pp 175–195
Quach K (2020) Researchers made an OpenAI GPT-3 medical chatbot as an experiment. It told a mock patient to kill themselves. https://www.utsa.edu/today/2020/07/story/chatbots-artificial-intelligence.html
Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union, L 119, 4 May 2016, pp 1–88
Renda A, Arroyo J, Fanni R, Laurer M, Sipiczki A, Yeung T (2021. Study to support an impact assessment of regulatory requirements for Artificial Intelligence in Europe: Final report, p. 203. Publications Office. https://doi.org/10.2759/523404
Reynoso R (2019) AI In Healthcare (+5 Ways It'S Used In 2020). Learn.G2.Com. https://learn.g2.com/ai-in-healthcare
Robert LP et al (2020) Designing fair AI for managing employees in organizations: a review, critique, and design agenda. Human-Comput Interaction 35(5–6):545–575
Rousseau A-L et al (2020) Doctor GPT-3: hype or reality? Nabla. https://www.nabla.com/blog/gpt-3/
Seth A (2019) A new paradigm to accommodate ethical foundations in the design and management of digital platforms. Working paper
Silverstone R (1999) Why study the media? SAGE Publications Ltd, London
Van Dijck J (2013) The culture of connectivity: a critical history of social media. Oxford University Press, Oxford
van Dijck J, Poell T, de Waal M (2018) The Platform Society: Public Values in a Connective World. Oxford University Press, Oxford