AI & SOCIETY (2025) 40:1485–1495
https://doi.org/10.1007/s00146-024-01972-6

NETWORK RESEARCH

Perceived responsibility in AI-supported medicine

S. Krügel[1] · J. Ammeling[1] · M. Aubreville[1] · A. Fritz[2] · A. Kießig[2] · Matthias Uhl[1]

Received: 8 January 2024 / Accepted: 6 May 2024 / Published online: 22 May 2024

Abstract

In a representative vignette study in Germany with 1,653 respondents, we investigated laypeople's attribution of moral responsibility in collaborative medical diagnosis. Specifically, we compare people's judgments in a setting in which physicians are supported by an AI-based recommender system to a setting in which they are supported by a human colleague. It turns out that people tend to attribute moral responsibility to the artificial agent, although this is traditionally considered a category mistake in normative ethics. This tendency is stronger when people believe that AI may become conscious at some point. In consequence, less responsibility is attributed to human agents in settings with hybrid diagnostic teams than in settings with human-only diagnostic teams. Our findings may have implications for behavior exhibited in contexts of collaborative medical decision making with AI-based as opposed to human recommenders because less responsibility is attributed to agents who have the mental capacity to care about outcomes.

Keywords AI-based recommender systems · Medical diagnosis · Collaborative intelligence · Ethics of AI · Responsibility gap

1 Introduction

Artificial intelligence (AI) increasingly supports diagnosticians like radiologists and pathologists in medical image recognition. Examples of its successful use include detection of kidney and lung disease, breast cancer and diabetes (Kaur et al. 2020). From an ethical perspective, decision support by AI in medical diagnosis seems desirable as there is ample evidence that AI may help to improve human diagnosis substantially (see also Grote and Berens 2019). Given the increasing number of diagnoses that a diagnostician has to make each day, recommender systems can help to reduce errors that are caused by fatigue or time pressure (see also Krupinski 2015). In certain contexts, it may be the collaborative intelligence of a human professional and an AI-based recommender system that features the best results (Bertram et al. 2021). This is the case, if the errors that humans and machines tend to make are less than perfectly positively correlated. With further advancements in AI, however, we may get to a point where AI will outperform humans to such a degree that collaboration with humans only dilutes overall performance.

Even if this point could be reached in the future, there may be an ethical case for limiting the role of AI to that of a recommender: the desire to attribute moral responsibility for moral evils. From a normative perspective, one may insist that the role of AI is limited to giving recommendations, because AI is not considered a moral agent who can be held morally responsible if something goes wrong. Several authors seem to take such a critical stance toward the delegation of moral tasks to autonomous systems and argue that only if the ultimate decisions are still taken by humans can we clearly attribute moral responsibility for a given wrong (or correct) outcome (Coeckelbergh 2021; Sparrow 2007). In a seminal paper, Matthias (2004) has introduced the notion of a "responsibility gap" in the context of learning machines. It captures the idea that the tendency to hold the manufacturer or operator of a machine responsible for the consequences of its operation may no longer apply for self-adapting machines because the manufacturers or operators can, as a matter of principle, not predict the machine's behavior anymore. For Matthias, holding manufacturers or operators morally responsible for something over which they

✉ Matthias Uhl
matthias.uhl@thi.de

[1] Faculty of Computer Science, Technische Hochschule Ingolstadt, Esplanade 10, 85049 Ingolstadt, Germany

[2] Faculty of Theology, Katholische Universität Eichstätt-Ingolstadt, Eichstätt, Germany

could not have sufficient control would be unjust. The usage of these machines then implies a responsibility gap. In parts of the literature, it is discussed whether this problem requires less individualistic notions of responsibility like distributed responsibility that may be more adequate for the information society (Floridi 2013; Isaacs 2017; Fritz et al. 2020).[1] List (2021) also notes that, for the first time in human history, freely operating AI systems that make high-stakes decisions will necessarily cause harm which leads to new ethical and regulatory challenges. He doubts that the complexity of these systems is reducible to human responsibility because the entirety of human responsibility may fail to do justice to the full amount of harm caused. In contrast to the more "fatalistic" attitude of Matthias (Santoni de Sio and Mecacci 2021), however, List (2021) argues that to avoid the creation of responsibility gaps, a concept of "AI responsibility," like "corporate responsibility," might be defensible. He argues that if we are prepared to consider group agents as moral agents, there is no in-principle barrier to view AI systems as qualified to be held responsible.

Tigard (2021) writes that responsibility is a dynamic enterprise filled with ambivalence and denies that there is a uniquely technology-based responsibility gap. One may take a more objective view toward people and exempt them from reactive attributes, while still seeing them as controllable and manageable. Analogous to people, "AI systems too are naturally exempted from our usual moral attitudes, but they can nonetheless be controlled, managed, manipulated, and trained" (Tigard 2021, p. 605). It is an open question whether Tigard is right in claiming that AI is excluded from "our usual moral attitudes" in the psychological sense. Put differently, it is an empirical question whether the attribution of moral responsibility to a medical recommender system—that may be considered a category mistake from a normative perspective by some and defensible by others—aligns with the psychological reality of laypeople's attribution of responsibility in complex socio-technical systems.

Investigating laypeople's perception of whether an AI can be held responsible is more than an academic exercise because the psychological realities of moral responsibility attribution may also co-determine the incentives of several agents in the healthcare system. There exists, for instance, empirical evidence that human agents may be able to deflect punishment from themselves by delegating tasks to machines (Feier et al. 2022; Krügel et al. 2023). This psychological effect is untouched by the normative idea that it is impossible to delegate responsibility to machines because delegators retain responsibility for the outcome resulting from their delegation (Di Nucci 2021). Anticipating the deflection of punishment by machines may incentivize delegation in morally critical cases.

Similarly, if a recommender system psychologically absorbs moral responsibility in case of a wrong diagnosis, this may reduce the moral burden perceived by the human agents, which may in turn increase their tolerance for wrong diagnosis. In this sense, distributed responsibility may feed back into the actions of human agents within the system (see also, for instance, Braun et al. 2020; Kempt and Nagel 2021; Kempt et al. 2022). Lang et al. (2023) argue that AI-induced responsibility gaps may be addressed if relevant stakeholders responsibilize these gaps by taking on moral responsibility for things that they are not, strictly speaking, blameworthy. In a similar vein, Kiener (2022) discusses the possibility to bridge AI's responsibility gap at will by people taking retrospective answerability, i.e., by humans making themselves morally answerable for harm caused by AI systems after this harm has occurred. The feasibility of these solutions may, at least partly, depend on whether it is supported by people's moral intuitions. After all, successful responsibilization or answerability may require that someone is held responsible by someone with whom this answer morally resonates. In this study, we therefore empirically tackle the question of how people ascribe moral responsibility in the case of a consequential medical diagnosis supported by an AI-based recommender system.

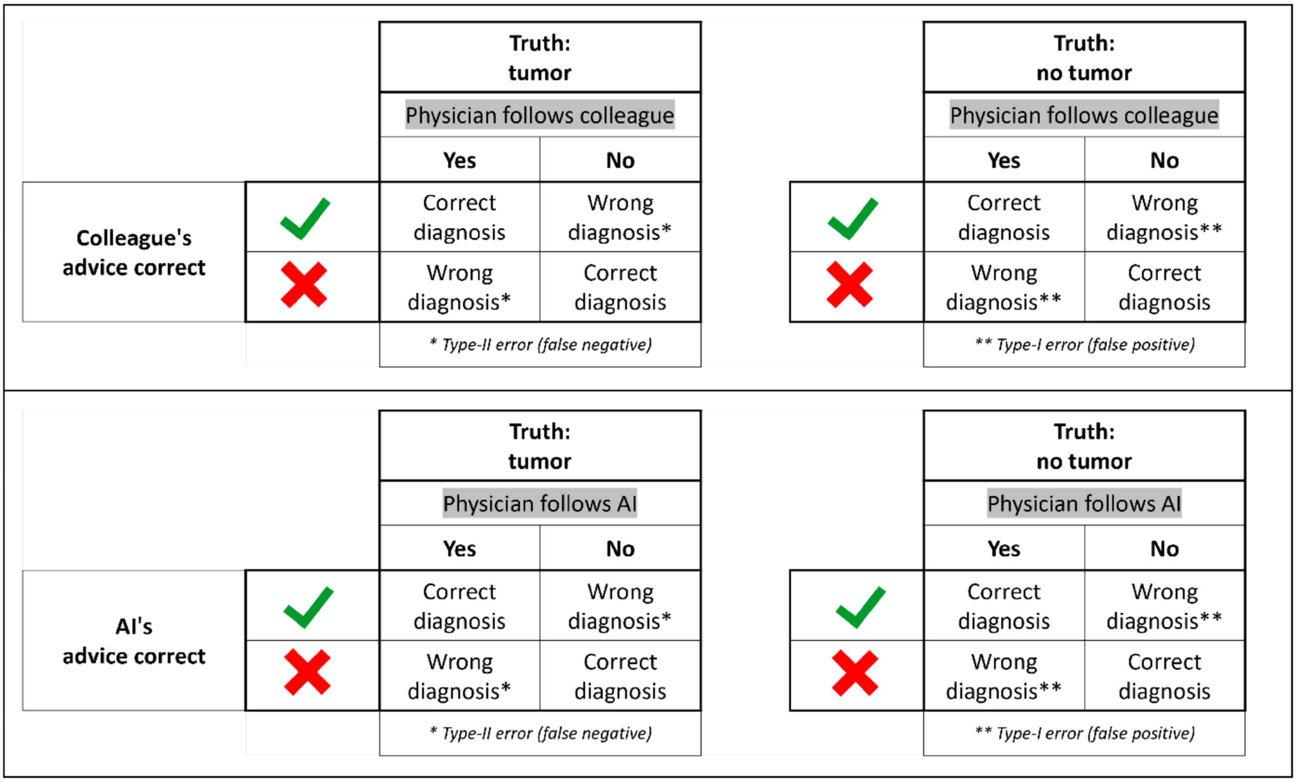
The paper proceeds as follows. In the second section, we describe the procedure and setup of our vignette experiment. In the third section, we present the experiment's results. In the fourth and final section, we discuss some implications of our findings.

2 Methods

2.1 Study design

We investigate the attribution of responsibilities for correct and incorrect medical diagnoses in a case of a collaborative tumor detection. Participants read a vignette that describes a situation in which a patient named Maria is being treated by a physician as part of her cancer screening. Our main interest is to compare the situation in which the physician uses an AI-based recommendation system for diagnosis with the situation in which the physician is advised by a human colleague. Overall, we manipulate four experimental dimensions in two variants each. We vary (1) the nature of the recommender (i.e., AI system or colleague), (2) the recommendation for diagnosis that the recommender gives (i.e., tumor or no tumor), (3) the truth content of the

[1] Against the mainstream, Danaher (2022) argues that techno-responsibility gaps may even be virtuous if humans can relieve themselves from tragic choices in moral dilemmas in which it is impossible to perfectly balance moral considerations by delegating them to AI agents.



		Truth: tumor	
		Physician follows colleague	
		Yes	No
Colleague's advice correct	✓	Correct diagnosis	Wrong diagnosis*
	✗	Wrong diagnosis*	Correct diagnosis
		* Type-II error (false negative)	

	Truth: no tumor	
	Physician follows colleague	
	Yes	No
✓	Correct diagnosis	Wrong diagnosis**
✗	Wrong diagnosis**	Correct diagnosis
	** Type-I error (false positive)	

		Truth: tumor	
		Physician follows AI	
		Yes	No
AI's advice correct	✓	Correct diagnosis	Wrong diagnosis*
	✗	Wrong diagnosis*	Correct diagnosis
		* Type-II error (false negative)	

	Truth: no tumor	
	Physician follows AI	
	Yes	No
✓	Correct diagnosis	Wrong diagnosis**
✗	Wrong diagnosis**	Correct diagnosis
	** Type-I error (false positive)	

Fig. 1 Scenarios in a situation of medical diagnosis with AI-based and human recommender

recommendation (i.e., true or false), and (4) the impact of the recommendation on the physician (i.e., follows or does not follow recommendation). These variations yield a total of 2 (recommender) × 2 (recommendation) × 2 (truth content) × 2 (impact) = 16 different scenarios.

The 16 different scenarios are illustrated in Fig. 1. A correct diagnosis emerges from the physician following a correct or ignoring a wrong advice of the AI or colleague and diagnosing the presence of an actually present tumor or the absence of an actually absent tumor. A wrong diagnosis emerges from the physician following a wrong or ignoring a correct advice of the AI or colleague and diagnosing the absence of an actually present tumor (type-II error or false negative) or the presence of an actually absent tumor (type-I error or false positive).

For means of illustration, we reprint here the respective vignette for the case that an AI-based recommendation system [a colleague] overlooks an existing tumor and where the physician follows its faulty advice.

> *Maria was examined by a* physician *as part of her cancer screening. The* physician *analyzed a medical image of her breast to see if a tumor was visible. During the examination, the* physician *was assisted by an artificial intelligence [a colleague]. The artificial intelligence [colleague] identified no tumor in Maria. The* physician *had to decide whether to follow the recommendation of the artificial intelligence [colleague]. The recommendation of the artificial intelligence [colleague] was in fact wrong because Maria does have a tumor disease. The* physician *decided to follow the wrong recommendation of the artificial intelligence [colleague], whereupon Maria did not undergo the necessary intervention.*

As a key dependent variable, participants then had to indicate their agreement with the statement "The physician bears the moral responsibility for the described wrong treatment of Maria" on a Likert scale, ranging from 0 ("no responsibility") to 6 ("full responsibility"). On a separate screen, we measured the attribution of moral responsibility to other agents who are explicitly or implicitly involved in the situation. These other agents are (1) the human colleague or the AI-based recommender system, (2) the hospital's management, (3) the physician's instructors,[2] (4) the colleague's instructors or the recommender system's developers, and (5) the patient herself. Each variable is measured in

[2] The word "instructors" is a translation from the German word "Ausbilder" which was used in the original wording to either refer to the academic teachers of the physician or to senior professionals who instructed him or her in medical practice.

the same way as the key dependent variable on a Likert scale by using a separate question for each variable. In addition to these main variables of interest, we asked participants whether they believe machines can develop consciousness.

2.2 Procedure

We conducted a representative online survey in Germany via the service provider Cint (http://de.cint.com).[3] Participation was estimated to take about 10 min. The cost per participant was €1.90. The study was pre-registered as #102265 at https://aspredicted.org. In line with our pre-registration, we aimed at a total of 1,600 respondents (i.e., approximately 100 respondents per scenario). At the beginning of the survey, participants had to declare their consent to participate. Subsequently, one randomly chosen scenario out of our 16 scenarios was described to each participant. To ensure that the participants had read and understood the respective scenario, they had to answer two control questions. Only participants who answered these two control questions correctly qualified to take part in the rest of the survey. Respondents were invited until our target size of 1,600 respondents was reached. In total, 2,376 respondents participated in the study, of whom 1,653 answered both control questions correctly. As pre-registered, the statistical analysis is based on these 1,653 responses.

The experiment was ethically approved by the German Association for Experimental Economic Research (https://gfew.de/en) under reference 8IZwRRbo. The investigation was conducted according to the principles expressed in the Declaration of Helsinki. Participants could terminate the survey at any time. Table 1 summarizes the demographic characteristics of our sample for the scenarios in which a colleague was involved in the diagnosis and in which an AI was involved. Respondents in both kinds of scenarios do not differ in terms of age ($p = 0.580$, unpaired t test), in terms of gender ($p = 0.575$, Chi-squared test) and the job sector in which they work ($p = 0.326$, Chi-squared test). This suggests that the random assignment of our respondents to the scenarios involving a human recommender and those involving an AI recommender proved successful.

Table 1 Demographic sample characteristics

	Recommender	
	Colleague	**AI**
Observations	816	837
Gender		
Male	395 (48.4%)	425 (50.8%)
Female	418 (51.2%)	410 (49.0%)
Other	3 (0.4%)	2 (0.2%)
Age	43.67 (14.01)	44.06 (14.35)
Job sector		
Agriculture, forestry and animal husbandry	7 (0.9%)	10 (1.2%)
Technology, computer science, engineering	86 (10.5%)	103 (12.3%)
Commerce, trade, tourism, administration	208 (25.5%)	237 (28.3%)
Social affairs, teaching	70 (8.6%)	52 (6.2%)
Health, medicine	82 (10.0%)	75 (9.0%)
Academic research	11 (1.3%)	7 (0.8%)
Other sector	211 (25.9%)	202 (24.1%)
Currently not in a job	141 (17.3%)	151 (18.0%)

For "Age", the numbers represent the means and, in parentheses, standard deviations. For all other variables, the numbers represent the number of observations in the corresponding category and, in parentheses, their proportion of all observations in the respective experimental condition (i.e., "Colleague" or "AI")

3 Results

Under the assumption that responsibility ought to be attributed only to human agents and not to AI agents, we first examine whether the total moral responsibility assigned to all human agents involved in the described situation is smaller if the physician is supported by an AI system rather than a colleague. If this is true, moral responsibility that could be meaningfully attributed in case of a collaborative situation of diagnosis with physician and human recommender evaporates in the collaborative situation with physician and AI-based recommender. As the left panel of Fig. 2 illustrates, our results suggest that this is in fact the case: the total responsibility attributed to all human agents is significantly smaller in the scenario in which an AI system is involved in the collaborative diagnosis than in the scenario in which a colleague is involved (16.30 vs. 18.18, $p < 0.001$ in case of a correct diagnosis and 13.70 vs. 15.77, $p < 0.001$ in case of a wrong diagnosis, based on unpaired t tests). The gap can be explained by the fact that part of the total responsibility attributed to all agents is assigned to the AI-based recommender system (see Fig. 3). If we take into account the assigned responsibility to this AI-based system, the total moral responsibility ascribed to all agents involved in a setting of a collaborative diagnosis that is either correct or incorrect is strikingly similar between the scenario in which a colleague advises the treating physician and in which an AI system gives this advice (18.18 vs. 18.01, $p = 0.675$ in case of a correct diagnosis and 15.77 vs. 15.56, $p = 0.598$ in case of a wrong diagnosis, based on unpaired t tests). As mentioned above, however, according to the general opinion

[3] Data and the full set of the vignettes used will be made available upon reasonable request.

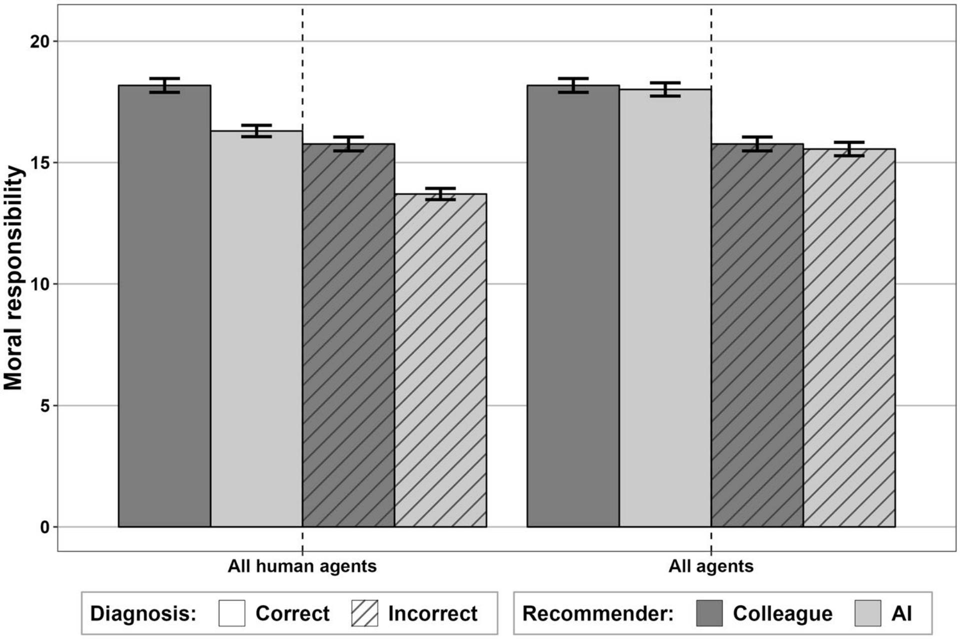


Fig. 2 Total moral responsibility assigned to all human agents and to all agents. Bars indicate means, and error bars indicate standard errors of the mean

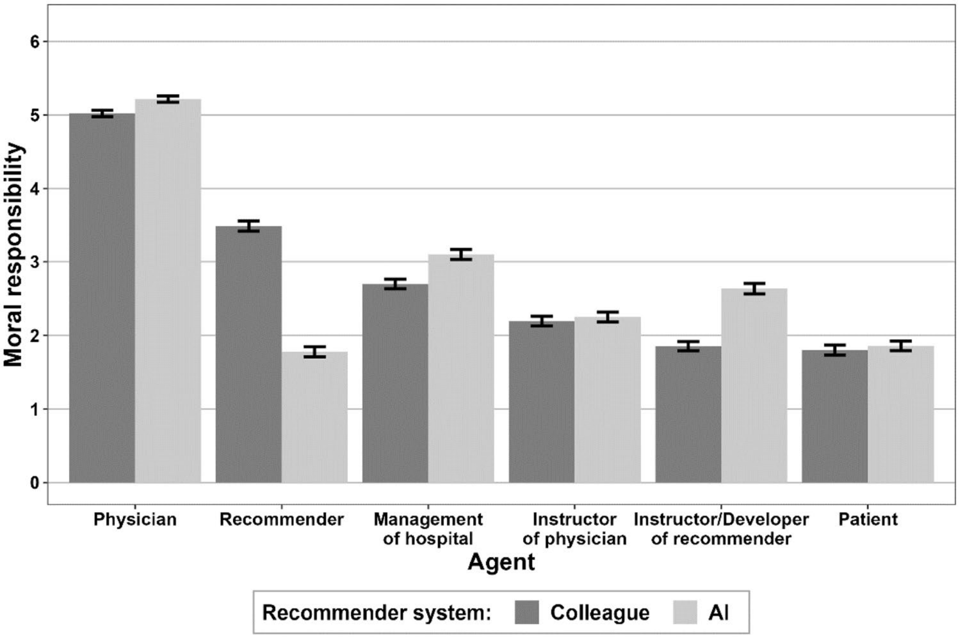


Fig. 3 Moral responsibility assigned to the respective agent. Bars indicate means, and error bars indicate standard errors of the mean

in the ethics of technology, moral responsibility cannot be meaningfully attributed to an artificial agent for lack of moral agency.

Figure 3 provides a more detailed overview of the ascription of responsibility to the attending physician and the other agents explicitly or implicitly involved in the situation. For ease of exposition, this figure does not (yet) differentiate between the level of moral responsibility ascribed in the cases of a correct and incorrect diagnosis. Notice that “physician”, “management of hospital”, “instructor of physician” and “patient” were identical between the vignettes in which a colleague was giving the recommendation and in which an AI was giving the recommendation. Whether responsibility could be attributed to a human or an AI-based “recommender” and whether it could be attributed to the “instructor” or the “developer” of the “recommender” depended on whether the physician in the respective scenario was advised by a colleague or an AI.

Regarding the treating physician, it turns out that slightly more moral responsibility was ascribed in a collaborative setting in which the physician is advised by an AI system than in one in which the physician is advised

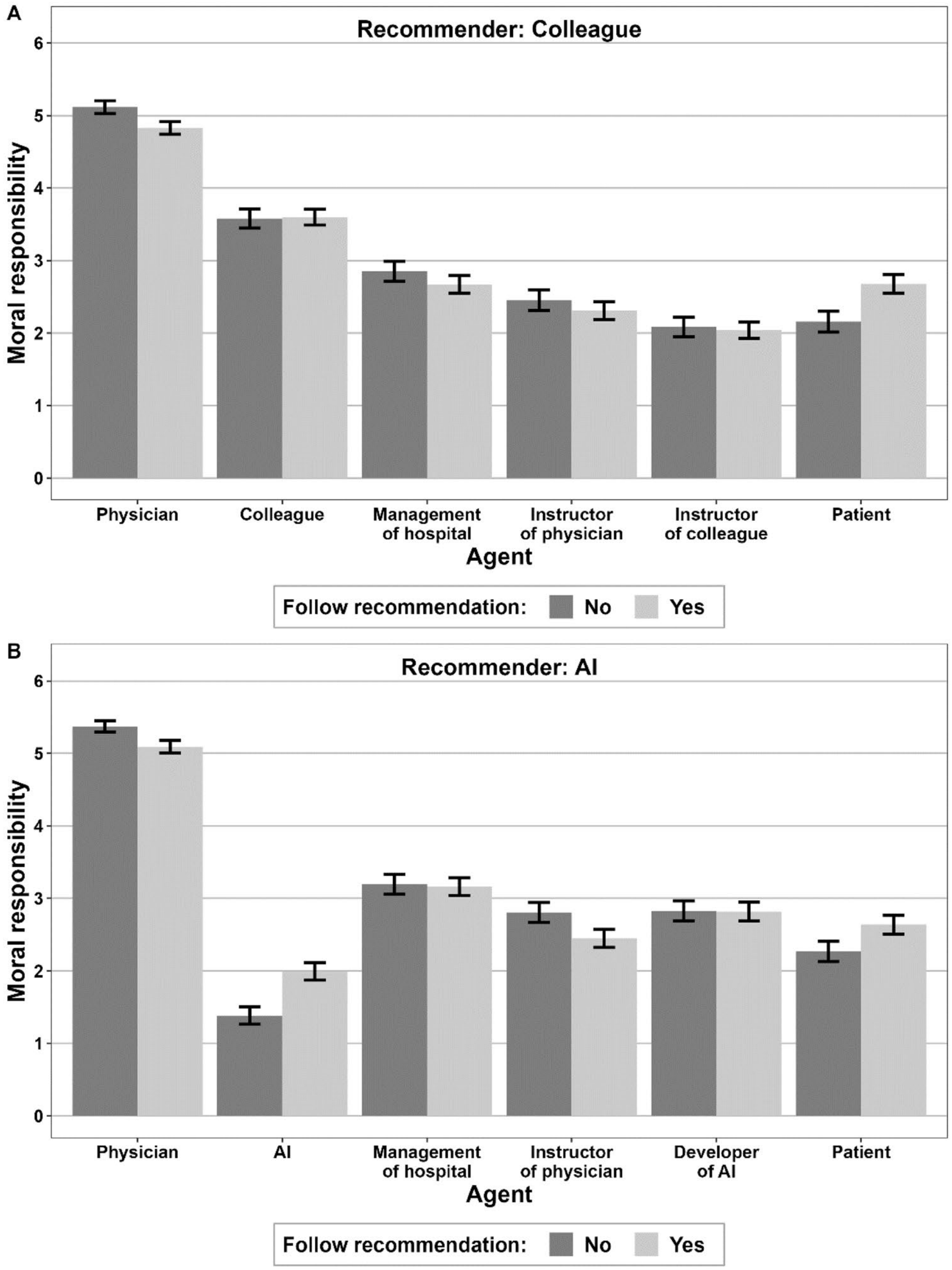


Fig. 4 Moral responsibility assigned to the respective agent in case of a correct diagnosis. Bars indicate means, and error bars indicate standard errors of the mean

by a colleague (5.22 vs. 5.02, $p = 0.002$, unpaired t test). The advising AI, in turn, attracted clearly less responsibility than the advising colleague (1.78 vs. 3.49, $p < 0.001$, unpaired t test). It is, however, noteworthy that the ascription of moral responsibility to the AI was not zero, but people actually attributed to it a considerable level of moral responsibility. In the scenario with the AI system acting as recommender, the hospital management was found to bear a greater level of responsibility than in case of a human as recommender (3.10 vs. 2.70, $p < 0.001$, unpaired t test). Moreover, the developers of the AI-based recommender attracted more responsibility than the instructors of the human recommender (2.63 vs. 1.86, $p < 0.001$, unpaired t test).

Figures 4 and 5 depict the ascriptions of moral responsibility to the involved agents in case of a correct and incorrect diagnosis. In addition, these figures distinguish whether or not the treating physician followed the recommendation of the colleague or AI. In case of a correct diagnosis, the ascription of responsibility to the various agents shows basically the same pattern as in Fig. 3. If the recommendation is based on an AI system instead of a colleague, the attending physician gets slightly more responsibility as well as the management of the hospital. The AI system itself is also

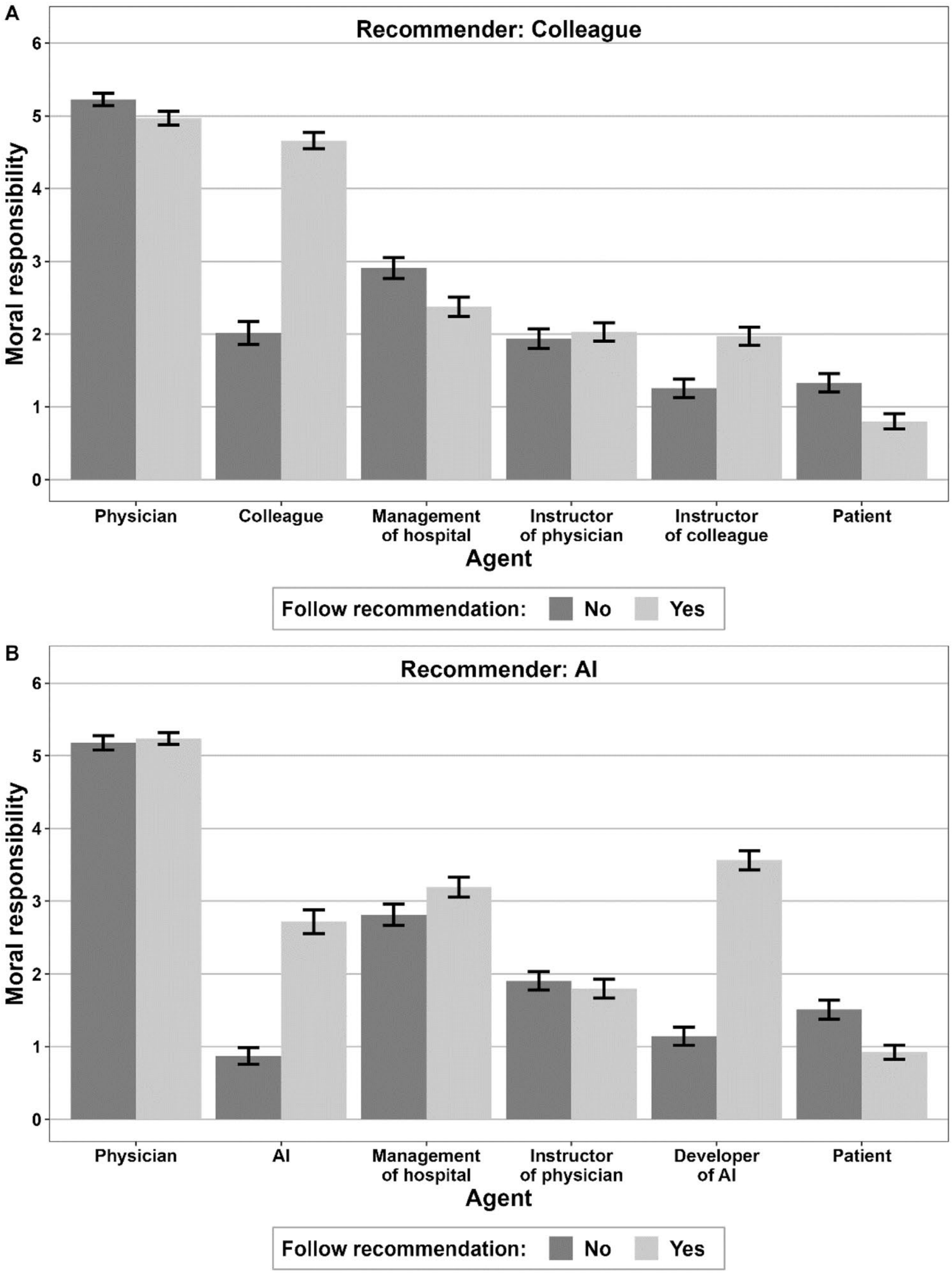


Fig. 5 Moral responsibility assigned to the respective agent in case of a wrong diagnosis. Bars indicate means, and error bars indicate standard errors of the mean

ascribed responsibility in case of a correct diagnosis, but substantially less than the advising colleague. Conversely, the developers of the AI system are ascribed considerably more responsibility than the instructors of the colleague. Beyond these differences between a colleague and an AI system as recommender, our subjects seem to be virtually indifferent to how a correct diagnosis came about. Whether the treating physician followed a correct recommendation of the colleague or AI, or overruled an incorrect recommendation, the attribution of responsibility to the individual agents remains the same. The only exceptions are the treating physician, who is assigned slightly more (positive) responsibility if he or she overruled an incorrect recommendation (5.12 vs. 4.83, $p = 0.021$ in case of the colleague as recommender and 5.37 vs. 5.09, $p = 0.018$ in case of the AI system as recommender, based on unpaired *t* test) and the AI system, which is assigned more responsibility if the recommendation was correct as opposed to incorrect (1.99 vs. 1.38, $p < 0.001$, unpaired *t* test).

In case of an incorrect diagnosis, it makes a huge difference to our subjects in terms of the allocation of responsibility to the involved agents whether the treating physician followed an incorrect recommendation or overruled a correct recommendation (see Fig. 5). For the treating physician himself or herself, however, there is almost no difference. He or she remains primarily responsible from our subjects' point of view. Only if the treating physician mistakenly overruled a correct recommendation of a colleague, the attribution of responsibility to him or her is slightly higher (5.23 vs 4.97, $p = 0.046$, unpaired *t* test). In this case, the management of the hospital is attributed slightly more responsibility as well (2.91 vs. 2.37, $p = 0.007$, unpaired *t* test). If the treating physician follows an incorrect recommendation of a colleague, this colleague is held substantially more responsible (4.66 vs. 2.02, $p < 0.001$, unpaired *t* test). From the subjects' point of view, the consulting colleague is almost equally responsible for the wrong diagnosis as the treating physician (4.66 vs. 4.97, $p = 0.037$, unpaired *t* test). Also, the colleague's instructors are considered slightly more responsible when the recommendation was incorrect than when it was correct (1.97 vs. 1.25, $p < 0.001$, unpaired *t* test).

If the treating physician was advised by an AI system (see Fig. 5b), this system is held substantially more responsible if the treating physician followed an incorrect recommendation than if the treating physician mistakenly overruled a correct recommendation (2.72 vs. 0.87, $p < 0.001$, unpaired *t* test). This effect is even more pronounced for the developers of the AI system (3.56 vs. 1.14, $p < 0.001$, unpaired *t* test). From our subjects' point of view, the developers of the AI system are the main culprits besides the treating physician when an incorrect diagnosis is made based on an incorrect recommendation. The AI system itself receives substantial responsibility as well, but significantly less than the consulting colleague (2.72 vs. 4.66, $p < 0.001$, unpaired *t* test). In contrast, the developers of the AI system are assigned significantly more responsibility than the instructors of the colleague (3.56 vs. 1.97, $p < 0.001$, unpaired *t* test). The management of the hospital also bears more moral responsibility when the incorrect recommendation comes from an AI system instead of a colleague (3.19 vs. 2.37, $p < 0.001$, unpaired *t* test).

If the diagnosis is incorrect and the treating physician was misled by an AI system instead of a colleague, responsibility is distributed differently among the involved agents from our subjects' point of view. The question is how this affects possible responsibility gaps in hybrid diagnostic teams. Figure 6 shows the total responsibility assigned to the involved agents in cases of a correct and incorrect diagnosis and depending on whether the physician followed the recommendation of the colleague or the AI system. If we only consider the total responsibility attributed to human agents, a responsibility gap occurs for hybrid diagnostic teams in all cases compared to human-only diagnostic teams. No matter whether the diagnosis is correct or incorrect and no matter how it came about, the total responsibility attributed to all human agents is significantly smaller for hybrid diagnostic teams ($p \leq 0.001$ in all four comparisons based on unpaired *t* test) and the responsibility gap is virtually the same in all cases. If we take into account the attributed responsibility to the AI-based recommender system, the responsibility gap disappears, except for the case of an incorrect diagnosis where the treating physician has overruled the recommendation. In the latter case, a responsibility gap seems to exist for hybrid diagnostic teams compared to human-only diagnostic teams even when the attributed responsibility to artificial agents is taken into account (13.43 vs. 14.67, $p = 0.030$, unpaired *t* test).

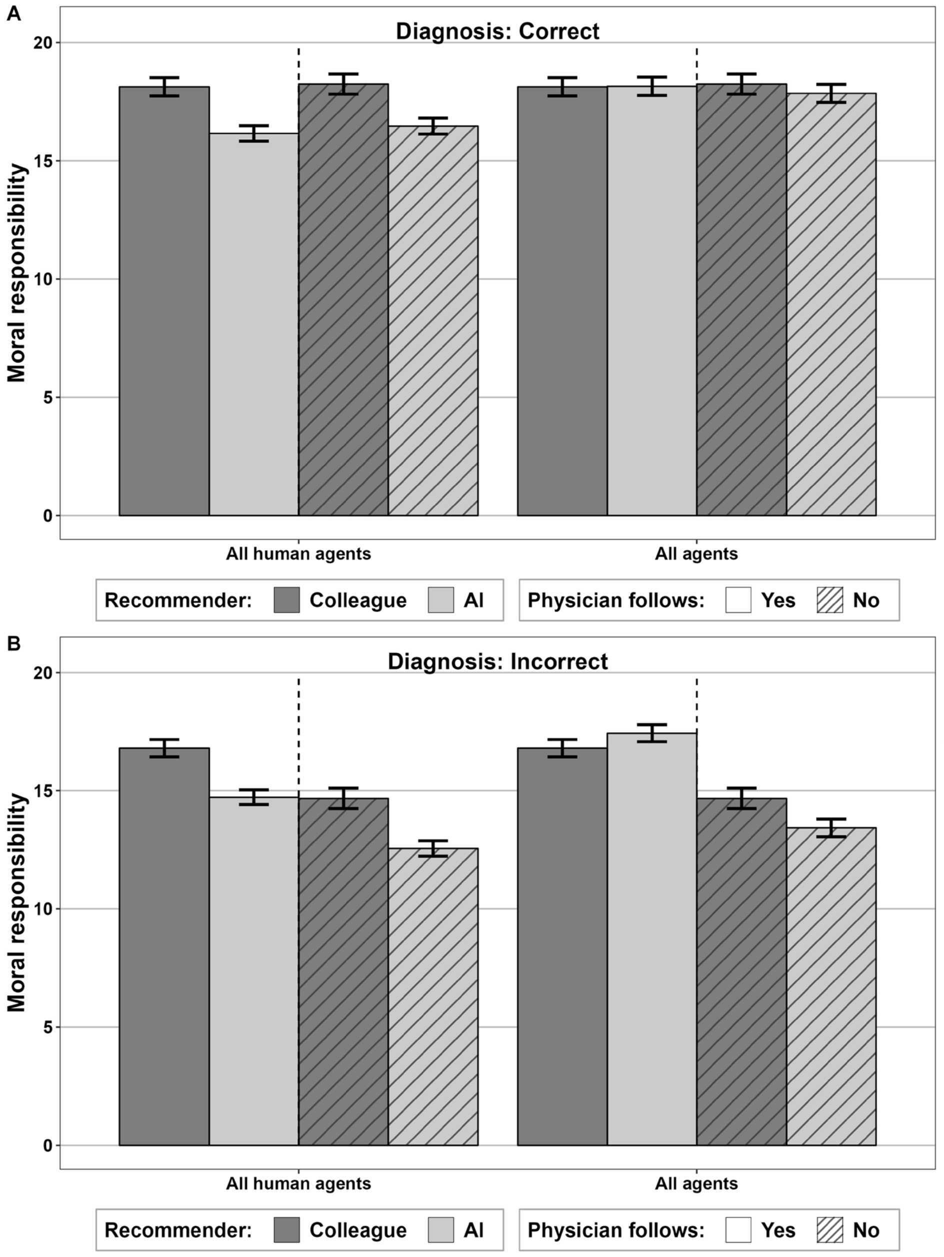
Additionally, we note that the error type was almost irrelevant for our subjects when an incorrect diagnosis was made. Whether a tumor was incorrectly diagnosed (false positive, type-I-error) or a tumor was missed (false negative, type-II-error) did not make a difference in the subjects' attribution of responsibility. The treating physician, the consulting colleague or AI system, the physician's instructors, the colleague's instructors or AI-system developers, and the patient all carry the same moral responsibility in case of a wrong diagnosis, whether the error is of type 1 or type 2 ($p > 0.2$ in each of these comparisons between attributed responsibility to each agent in cases of errors of type 1 and 2). Only the management of the hospital bears slightly more responsibility in case of a false positive diagnosis than in case of a false negative diagnosis when the treating physician was advised by a colleague (2.87 vs. 2.37, $p = 0.011$, unpaired *t* test). The indifference of error types in attributing responsibility for misdiagnosis to the involved agents is noteworthy because one might expect that the two types of errors are valued very differently.

Finally, one of the questions asked at the end of our experiment was whether participants believe that machines may at one point develop consciousness. People could agree to the respective statement on a Likert scale ranging from 0 ("do not agree at all") to 6 ("fully agree"). Figure 7 illustrates the correlation between increasing levels of agreement with this statement and the degree of moral responsibility ascribed to the respective recommender. Notice that the stronger the participants agree with the belief that machines may develop consciousness, the more do they attribute moral responsibility to the AI-based recommender system in the respective scenarios ($p < 0.001$). In contrast to this, the moral responsibility attributed to the human recommender is not correlated with the level of agreement that people express with respect to the belief that machines may develop consciousness ($p = 0.851$).

4 Discussion and conclusion

The future use of AI in medical diagnoses promises to reduce errors on the one hand and to cushion the increasing workload of physicians on the other. The use of AI is

Fig. 6 Total moral responsibility in case of a correct and wrong diagnosis. Bars indicate means, and error bars indicate standard errors of the mean



predominantly seen in the form of recommendation systems or possibly as a substitute in human-only medical teams. AI supports and the final diagnosis remains the responsibility of the physician. It is not clear, however, whether this conceptual determination corresponds with psychological reality because the perception of physicians and doctor–patient relationships may be affected. An important question that the emergence of hybrid diagnostic teams raises is therefore to what extent the use of AI changes ascriptions of moral responsibility. We surveyed a representative sample of (potential) patients in Germany on this issue.

In the view of our respondents, the main responsibility in medical diagnoses remains with the treating physician, regardless of whether he or she is advised by an AI system or a colleague. In human-only medical teams, it is mainly the consulting colleague and the hospital's management who are morally responsible for a diagnosis next to the treating physician. In hybrid diagnostic teams, the responsibility of the artificial recommender system is smaller than that of a consulting colleague, and, instead, the management of the hospital and the developers of the recommender system are regarded as key players next to the treating physician by our subjects. Nonetheless, the AI-based recommender system is attributed a significant moral responsibility that cannot be ignored. If the diagnosis in the respective team is correct, our subjects do not

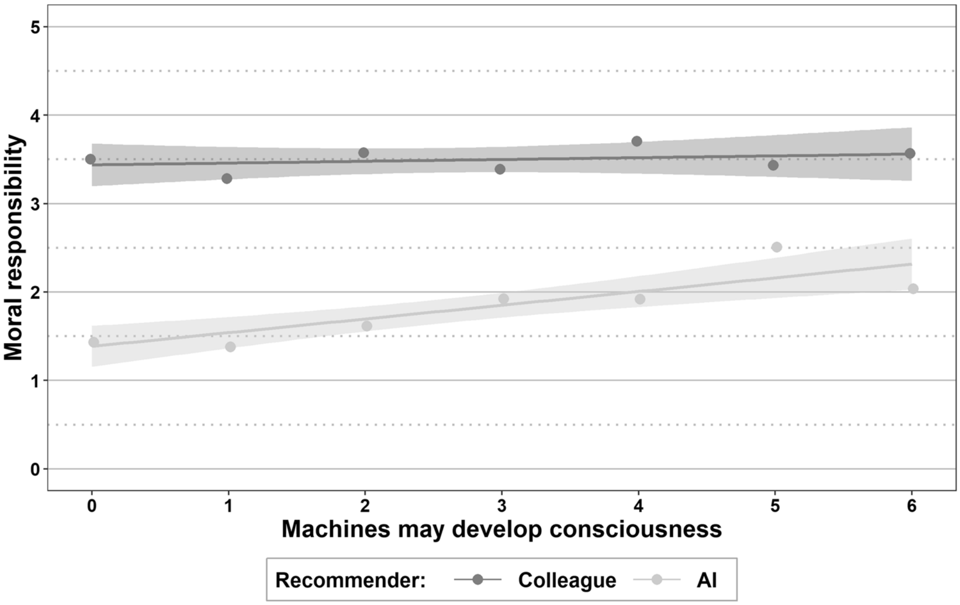


Fig. 7 Moral responsibility ascribed to recommender depending on agreement with statement that machines may develop consciousness

seem to care much about how the diagnosis came about. If the diagnosis is incorrect, however, the ascription of moral responsibility depends strongly on the diagnostic process. If the treating physician follows a (wrong) recommendation of the consulting colleague, the latter is almost as responsible as the physician, in the eyes of our subjects. If the treating physician follows a wrong recommendation of an AI-based recommender system, it is mainly the developers of the system and the management of the hospital in addition to the physician, but also the system itself, which is held responsible from our subjects' point of view. Interestingly, our subjects do not distinguish between the error types of incorrect diagnoses when allocating moral responsibility.

Some ethicists insist that AI is no moral agent and therefore cannot carry moral responsibility for the outcomes that it causally co-determines. The moral intuitions of laypeople, however, seem to differ from this normative premise. In a setting of medical diagnosis, respondents in our representative German sample factually ascribed responsibility to an AI-based recommender system to a non-negligible degree. This finding is in line with the result from a recent study which shows that people hold AI-powered car warning systems co-responsible for outcomes of actions taken by drivers who rely on the system's advice, even though they consider these AI agents as mere tools (Longin et al. 2023). Interestingly, in our study, the more people agree with the idea that machines may at some point in the future develop consciousness, the stronger do they attribute responsibility to the AI system. This leads to the overall effect that less responsibility is attributed to human agents in settings with hybrid diagnostic teams than in settings with human-only diagnostic teams.

As elaborated in the introduction, List (2021) argues that a responsibility gap caused by the use of AI is either fatalistically accepted or pragmatically avoided by the attribution of AI responsibility. He sees no reason to reject the idea that AI could, at least in principle, fulfil the necessary criteria to constitute a moral agent who is capable of being held responsible. It seems that this idea also resonates with the moral intuitions of the participants in our study, especially if they believe in the potential of an AI system developing consciousness. In this sense, those who are potentially affected may not *perceive* a responsibility gap by the usage of AI-based recommenders in medical diagnosis. Thus, if the responsibility gap is conceptualized in perceptual terms, it may not exist here, or it may at least be much less pronounced. Only in the case of an incorrect diagnosis, in which the treating physician disregarded a correct recommendation of the AI system, a perceptual responsibility gap was present in our data even if we included the attributed moral responsibility to the artificial agent.

If one objects to List's position on "AI responsibility", then responsibility gaps consist in the erosion of responsibility ascribed to entities that normatively qualify as full moral agents, i.e., humans. Among laypeople (i.e., the potential patients), such a gap was indeed present in our data if the physician's recommender was artificial instead of human. On a normative level, one might ask whether responsibility gaps in laypeople's perceptions are problematic. After all, laypeople's perceptions may be ignored in discussions of philosophical concepts. From an empirical perspective, however, it cannot be ruled out that the factual erosion of responsibility ascribed to human agents may alter the incentives of the actors in the healthcare system. This would, in turn, have normative implications, if misdiagnoses are more easily tolerated because less blame will be attributed to agents who do have the mental capacity to care about medical outcomes. It has yet to be tested whether this kind of responsibility gap could be effectively bridged by human stakeholders taking on the responsibility or retrospectively answering for incorrect diagnoses for which they are not, strictly speaking, blameworthy as Lang et al. (2023) or Kiener (2022) suggest. This solution may be psychologically challenging for two reasons. First, as also acknowledged by Lang et al. (2023), because of the supererogatory nature of willingly taking on responsibility for acts for which one is not fully to blame. But also, second, because this assumption of responsibility must fall on fertile ground ethically and emotionally with those who are at the attributing end of the blame game.

Funding Open Access funding enabled and organized by Projekt DEAL. Funding was provided by Bayerisches Forschungsinstitut für Digitale Transformation.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Bertram CA, Aubreville M et al (2021) Computer-assisted mitotic count using a deep learning-based algorithm improves interobserver reproducibility and accuracy. Vet Pathol 59(2):211–226

Braun M, Hummel P, Beck S, Dabrock P (2020) Primer on an ethics of AI-based decision support systems in the clinic. J Med Ethics 47(12):e3–e3

Coeckelbergh M (2021) Narrative responsibility and artificial intelligence: How AI challenges human responsibility and sense-making. AI & Soc 2021:1–14

Danaher J (2022) Tragic choices and the virtue of techno-responsibility gaps. Philos Technol 35:26

Di Nucci E (2021) The control paradox: from AI to populism. Rowman & Littlefield, Lanham

Feier T, Gogoll J, Uhl M (2022) Hiding behind machines: artificial agents may help to evade punishment. Sci Eng Ethics 28(2):19

Floridi L (2013) Distributed morality in an information society. Sci Eng Ethics 19:727–743

Fritz A, Brandt W, Gimpel H, Bayer S (2020) Moral agency without responsibility? Analysis of three ethical models of human-computer interaction in times of artificial intelligence (AI). De Ethica 6(1):3–22

Grote T, Berens P (2019) On the ethics of algorithmic decision-making in healthcare. J Med Ethics 46:205–211

Isaacs T (2017) Kollektive Verantwortung. In: Heidbrink L, Langbehn C, Loh J (eds) Handbuch Verantwortung. Springer, Wiesbaden, pp 453–475

Kaur S, Singla J et al (2020) Medical diagnostic systems using artificial intelligence (AI) algorithms: principles and perspectives. IEEE Access 8:228049–228069

Kempt H, Nagel SK (2021) Responsibility, second opinions and peer-disagreement: ethical and epistemological challenges of using AI in clinical diagnosis contexts. J Med Ethics 48(4):222–229

Kempt H, Heilinger J-C, Nagel SN (2022) Relative explainability and double standards in medical decision-making. Should medical AI be subjected to higher standards in medical decision-making than doctors? Ethics Inf Technol 24(2):20

Kiener M (2022) Can we Bridge AI's responsibility gap at Will? Ethical Theory Moral Pract 25:575–593

Krügel S, Ostermaier A, Uhl M (2023) Algorithms as partners in crime: a lesson in ethics by design. Comput Hum Behav 138:107483

Krupinski EA (2015) Improving patient care through medical image perception research. Policy Insights Behav Brain Sci 2(1):74–80

Lang BH, Nyholm S, Blumenthal-Barby J (2023) Responsibility gaps and black box healthcare AI: shared responsibilization as a solution. Digit Soc 2(3):52

List C (2021) Group agency and artificial intelligence. Philos Technol 24:1213–1242

Longin L, Bahrami B, Deroy O (2023) Intelligence brings responsibility—even smart AI assistants are held responsible. iScience 26(8):107494

Matthias A (2004) The responsibility gap: ascribing responsibility for the actions of learning automata. Ethics Inf Technol 6:175–183

Santoni de Sio F, Mecacci G (2021) Four responsibility gaps with artificial intelligence: why they matter and how to address them. Philos Technol 34:1057–1084

Sparrow R (2007) Killer robots. J Appl Philos 24(1):62–77

Tigard DW (2021) There is no techno-responsibility gap. Philos Technol 34:589–607