Journal of Nuclear Medicine, published on October 2, 2020 as doi:10.2967/jnumed.120.256032

Title: *When Does Physician Use of AI Increase Liability?*

Running title: *Liability for Physician Use of AI*

Authors: Kevin Tobia,[1,2]* Aileen Nielsen[2]* & Alexander Stremitzer [2]

1 Georgetown University Law Center
2 ETH Zurich, Center for Law & Economics
* These authors contributed equally to the study

Corresponding Author: Kevin Tobia, 600 New Jersey Avenue, Washington D.C., 20001, 201-572-3407, kevin.tobia@georgetown.edu

Funding/Support: This study was funded by ETH Zurich and the ETH Fellows Program.

ABSTRACT

Rationale: An increasing number of automated and artificially intelligent (AI) systems make medical treatment recommendations, including "personalized" recommendations, which can deviate from standard care. Legal scholars argue that following such nonstandard treatment recommendations will increase liability in medical malpractice, undermining the use of potentially beneficial medical AI. However, such liability depends in part on lay judgments by jurors: When physicians use AI systems, in which circumstances would jurors hold physicians liable?

Methods: To determine potential jurors' judgments of liability, we conducted an online experimental study of a nationally representative sample of 2,000 U.S. adults. Each participant read one of four scenarios in which an AI system provides a treatment recommendation to a physician. The scenarios varied the AI recommendation (standard or nonstandard care) and the physician's decision (to accept or reject that recommendation). Subsequently, the physician's decision caused a harm. Participants then assessed the physician's liability.

Results: Our results indicate that physicians who receive advice from an AI system to provide standard care can reduce the risk of liability by accepting, rather than rejecting, that advice, all else equal. However, when an AI system recommends nonstandard care, there is no similar shielding effect of rejecting that advice and so providing standard care.

Conclusion: The tort law system is unlikely to undermine the use of AI precision medicine tools and may even encourage the use of these tools.

Keywords: artificial intelligence, liability, precision medicine

INTRODUCTION

Imagine a woman has recently been diagnosed with ovarian cancer. To help determine the dosage of a chemotherapy drug, the treating hospital has adopted routine use of an artificially intelligent (AI) precision medicine tool. The AI advises, based on the patient's file, that a *nonstandard* dosage is most likely to succeed. But what if something goes wrong as a result of the treatment? Will the physician be judged harshly for accepting unorthodox treatment advice from a computer? Or might the physician be judged even more harshly for rejecting advice from a state-of-the-art tool?

The woman's story is a hypothetical example from an important recent paper on AI in medicine (*1*). But this story may not remain a hypothetical for long. Recent advances in AI medical technology make possible a wide range of "personalized" medical recommendation tools, some of which have achieved regulatory approval and are increasingly adopted by medical providers (*2*).

However, despite the promise of these AI medical systems to improve patient outcomes, legal scholars have cautioned that tort law may create a substantial legal barrier to physicians' uptake of AI recommendations: Accepting certain AI recommendations may increase physicians' risk of liability in medical malpractice (*1*). In particular, given tort law's privileging of "standard care," physicians who accept a personalized AI recommendation to provide nonstandard care would increase their risk of medical malpractice liability.

The purpose of this investigation is to contribute empirical evidence bearing on these questions: In which circumstances are physicians using artificially intelligent systems more likely to be found liable, and how can physicians reduce their potential liability?

The answer to this question depends, in large part, on ordinary people's judgments. Liability for medical malpractice turns on lay jurors' assessment of the "reasonableness" of medical treatment (*3*). A growing number of states rely on the "reasonable physician standard" (*4*). For example, to determine whether a physician's neglecting a patient who was vomiting blood is medical malpractice, the court asks whether "a layperson could infer that a reasonable physician, acting with the skill of other physicians in the community, would not neglect a patient vomiting blood" (*5*). If litigation arises for a physician's use of AI, the same reasonableness standard would apply.

Of course, only a fraction of medical malpractice lawsuits reach a jury—many more settle (*6*). But even parties who ultimately settle their medical malpractice claims benefit from knowledge about the likely jury outcome *if* trial ensued. For those, the results here provide evidence about the "shadow of the law"; the likely outcome of the court proceedings is an important input into settlement negotiations.

MATERIALS AND METHODS

Study Population, Design, and Setting

The study was conducted in March 2020. Participants were recruited through Lucid (luc.id) in the form of a nationally representative sample as stratified by age, race, and gender of two-thousand participants from the United States. Our sample size is based on a power analysis using G*Power. We ran power analyses to assess each of the pre-registered tests, with a power of .95, to detect effect sizes of a small size (f = .10). The calculations indicated that a sample of 1300 would be sufficient to assess each of the pre-registered tests at this level. We anticipated excluding up to 30% of participants (e.g. for failing comprehension check questions) and thus recruited 2000 participants.

Participants were presented with different versions of a medical AI scenario. To minimize researcher degrees of freedom, the scenarios closely followed Price et al. (*1*), who introduced these vignettes without knowledge of the hypothesis of this project. In each scenario, a physician reviewed an ovarian cancer patient's case file, which included routine input from a medical AI system called "Oncology-AI." Participants were told that the AI system had all of the relevant regulatory approvals and provided a chemotherapy drug dosage recommendation.

The study had a 2x2 between-subjects factorial design, varying the AI Recommendation (recommending standard or nonstandard treatment) and the physician's Decision (accepting or rejecting the AI recommendation). Thus, participants received one of four possible scenarios: A physician receives from an AI medical system either a standard treatment recommendation (900 mg every three weeks) *or* a nonstandard treatment recommendation (4,500 mg every three weeks). Then, the physician either accepts the recommendation *or* rejects it. See Figure 1. In all scenarios, the treatment choice causes a harm and participants evaluate these facts according to the same legal standard: was the decision one that could have been made by a “reasonable physician”?

[Figure 1 here]

The study was designed to adjudicate among four plausible models of lay judgment of legal liability, with each model resulting from a different combination of two factors: provision of standard care and adherence to the AI recommendation. These four models make very different predictions about the pattern of results across the four scenarios and what a physician should do to minimize liability (summarized in Figure 2).

1. Follow AI Recommendations. The first possibility is that lay jurors are more inclined to hold physicians liable for rejecting AI recommendations. Some legal scholars have suggested that such a model is likely in the future, as the use of AI precision medicine grows (*1*, *7*).

2. Follow Standard Care. A second possibility is that lay jurors are more inclined to hold physicians liable for providing nonstandard care, regardless of the AI recommendation. This reflects Price et al’s (*1*) presentation of current tort law.

3. Follow Both. A third possibility is that lay judgment is affected by both factors: AI recommendation and standard care. This model predicts a significant interaction between Recommendation and Decision and further predicts that the mean ratings for Standard-Accept will be greater than mean ratings for the other three treatments.

4. Discretionary Model. A final possibility is that neither factor plays a significant role in laypeople’s liability determinations.

[Figure 2 here]

Our hypothesis is that lay judgments of liability are driven by *both* whether the AI recommended the treatment *and* whether the treatment is standard. If the hypothesis is true, we expect to find the "Follow Both" pattern of results, given the experimental design (https://osf.io/zyejh/?view_only=b0158bcde6a64d0da3a977b44b0610ca).

The statistical test selected for our primary pre-registered hypothesis was a 2*2 ANOVA. For the additional pre-registered hypotheses, we employed two-sided t-tests to test for a significance difference and two one-sided t-tests to test for equivalence between the two conditions in which the two influential factors were in conflict (Nonstandard-Reject and Nonstandard-Accept). In the case of the two one-sided t-tests analysis, we pre-specified a medium effect size of 0.5. In all cases the a priori significance level was .05. Unless otherwise noted, analyses were conducted with Stata 15.1MP, which is developed by StataCorp. All authors participated in the data analysis, as well as Dr. Henry Kim, who provided technical research assistance.

Procedure

The full experimental protocols are available in the online supplementary materials (https://osf.io/zyejh/?view_only=b0158bcde6a64d0da3a977b44b0610ca) and in the Appendix Section I (*8*, *9*). A schematic showing the overall flow of the survey scenario text ("vignettes") is in Figure 3.

Participants were randomly assigned to one scenario (see Figure 1) and evaluated the reasonableness of the physician's decision on a Likert scale from 1-7: Was the physician's treatment decision on that "could have been made by a reasonable physician in similar circumstances." Higher scores indicated greater reasonableness and thus lower liability.

The experiment was run with approval from the ETH Institutional Review Board, and all participants provided written informed consent before participating in the experiment.

[Figure 3 here]

RESULTS

The study, exclusion criteria, analyses, and hypotheses were pre-registered with AsPredicted, and all pre-registered analyses are reported. The initial response rate (responses to Lucid's emails) was ninety-seven percent; the conversation rate (finishing the study) was seventy-eight percent. Two-thousand and sixty participants completed the study, 693 were excluded for failing pre-registered comprehension checks and 11 for other reasons (e.g. taking the survey twice; see Appendix Section II for details). The results are robust to analyzing the data without exclusions (see https://osf.io/zyejh/?view_only=b0158bcde6a64d0da3a977b44b0610ca).

We conducted a 2(Recommendation: standard care, nonstandard care) * 2(Decision: accept, reject) ANOVA, treating reasonableness ratings as the dependent variable. As predicted, there was a main effect of Decision, $F(1, 1352) = 167.71$, $p < .0001$, $\eta_p^2 = .11$, and a significant Decision * Recommendation interaction, $F(1, 1352) = 51.68$, $p < .0001$, $\eta_p^2 = .037$. There was no significant effect of Recommendation, $F(1, 1352) < 1$, $p = 0.95$. (see Figure 4). Ratings were highest for Standard-Accept (*M* = 5.77, 95% CI (5.60, 5.95)), then Nonstandard-Accept (*M* = 5.09, 95% CI (4.91, 5.27)), then Nonstandard-Reject (*M* = 4.55, 95% CI (4.35, 4.75)), and finally Standard-Reject (*M* = 3.87, 95% CI (3.68, 4.07)). Table 1 presents pairwise comparisons among conditions.

[Figure 4 here]

[Table 1 here]

Following our pre-registration plan, we next evaluated several pairwise comparisons of interest (see Table 1 for statistical reporting). The mean reasonableness rating in Standard Accept was significantly higher than that in Standard Reject. It was also significantly higher than either of the Nonstandard conditions. We predicted that the difference between Nonstandard-Accept and Nonstandard-Reject would be small (less than d = .5). The analysis was consistent with the equivalence hypothesis with a pre-registered medium-sized effect (Cohen's d = 0.5), $t(596) = -2.46$, $p = .007$, 90% CI (.31, .77) . Ratings were significantly higher for Nonstandard-Accept compared to Nonstandard-Reject, but the size of this effect was small. All pairwise t-tests were to test pre-registered hypotheses except for the comparison of the Nonstandard Accept and Nonstandard Reject means to Standard Reject, which are post hoc comparisons.

The overall pattern of these results is most consistent with the "Follow Both" model, and we take this to suggest that lay jurors rely on both factors (AI recommendation, provision of standard care). However, one might wonder whether this pattern is a composite of the "Follow AI" and "Follow Standard Care" models. Perhaps some participants think that a physician ought to only follow the AI recommendation, while others think that a physician should only act according to what is considered standard care in reaching their liability assessment. Under such a "heterogeneous types" hypothesis, the "Follow Both" pattern reflects a mixture of those types in the subject pool.

However, the data do not support this "heterogeneous types" view. Across all four experimental conditions, distributions of the reasonableness ratings were unimodal and not bimodal (see https://osf.io/zyejh/?view_only=b0158bcde6a64d0da3a977b44b0610ca). The "heterogenous types" view would predict bimodal distributions when the two AI and standard care factors diverge (e.g. in "Nonstandard-Accept"). Moreover, in a series of pre-registered follow-up questions, participants ranked how important the factors "Follow AI recommendation" and "Providing standard care" were to their reasonableness assessment. Most (77%) rated the importance of both factors at or above the midpoint (https://osf.io/zyejh/?view_only=b0158bcde6a64d0da3a977b44b0610ca).

In addition to these main results, which reflect the analyses in the pre-registration, we also collected exploratory data on a number of other factors, including demographic data (the full battery is available at https://osf.io/zyejh/?view_only=b0158bcde6a64d0da3a977b44b0610ca). The main ANOVA findings (see Figure 3) are robust to including age, race, and gender as covariates (see Appendix Section III Table A1).

DISCUSSION

Overall, the results provide strong support for the “Follow Both” model of lay liability judgment. People use two different factors to evaluate physicians using medical AI systems: whether the treatment provided was standard and whether the physician followed the AI recommendation.

These results have important implications for physicians who seek to minimize tort liability. If a physician receives a standard care AI recommendation, there is a legal incentive to accept it. All else equal, participants judge accepting a standard care recommendation as more reasonable than rejecting it. On the other hand, if a physician receives a nonstandard AI recommendation, she does not necessarily make herself safer from liability by rejecting it.

Given that physicians who receive nonstandard advice are worse off in terms of liability than physicians who accept standard advice, healthcare institutions might consider whether to make AI systems available to physicians. However, the experimental scenarios studied here assume that an AI recommendation is *already* routinely offered. The study has nothing to say about the relative likelihood of liability for a physician who has not received advice from an AI system and therefore does not support any inference that healthcare institutions should avoid introducing AI systems. Additionally, those decisions will likely involve non-legal factors as well, such as the competitive pressure to maintain state-of-the-art facilities and their ability to set guidelines for the appropriate use of the AI system.

The study most directly examines laypeople as potential jurors, but it also sheds light on laypeople as potential patients. Important recent work in psychology shows laypeople are “algorithm averse” in other forecasting contexts, particularly when they see algorithms err (*10*). But this study’s results suggest that laypeople are not as strongly averse to physicians’ acceptance of precision medicine recommendations from an AI tool, even when the AI errs. The two scenarios rated most “reasonable” were also the two in which the algorithm’s diagnostic advice was wrong for the patient (Standard-Accept and Nonstandard-Accept). In the other two scenarios (Standard-Reject and Nonstandard-Reject), the physician rejected correct AI advice, and this was evaluated as more unreasonable. Together, these results suggest that patients may not exhibit strong algorithm aversion in such medical contexts.

Finally, the findings also speak to recent concerns about legal impediments to the use of AI precision medicine (*1*). Tort law may not impose as great a barrier to the uptake of AI medical system recommendations as is commonly assumed; in fact, it might even encourage the uptake of AI recommendations.

Moreover, we find the same Decision and Decision * Recommendation effects on ratings of whether *most physicians* in similar circumstances would have made the same treatment decision (See Appendix IV) (*11*). This suggests that the results extend to medical negligence standards focused more squarely on judgment of what care is common or customary. More broadly, these exploratory findings are consistent with prior research indicating that lay conceptions of *reasonableness* are affected by what seems common (See Appendix IV) (*11*). If this is right, we would predict that as AI use becomes more common, any tort law incentive to accept AI recommendations will only strengthen further. And we would predict this both for medical negligence standards centered on custom and those expressed more broadly in terms of "reasonableness."

Limitations

This study concerns judgments of liability, *given* that a harm has occurred. In practice, liability risk is determined by two factors: (factor a) the probability of liability given a harm (the focus of our study) *and* (factor b) the probability of a harm occurring at all. The first factor addresses whether tort law may be a barrier to AI use in medicine. The second factor, however, is best estimated by medical experts with rich knowledge of the specific context. For example, there could be a medical context in which a physician receives AI advice to provide standard care, but the physician is *extremely* confident that this is the wrong advice and that accepting it will harm the patient. Our findings suggest that the physician would have a degree of protection from the tort law system, which favors providing standard advice and following the AI's recommendation (factor a). However, if the probability of harm is sufficiently great (factor b), it could -- and should -- outweigh the tort law incentive to accept standard advice.

There are also limitations in the degree to which the study modeled the procedural elements of a real jury trial. For example, jurors would likely be presented with *expert testimony* concerning the use of AI precision medicine. Of course, jurors would normally hear expert testimony from each side, one expert that favored taking AI advice and one that disfavored it. Future work could assess whether there is any systematic effect of dueling expert testimony in these cases (e.g. perhaps laypeople generally tend to defer to pro-AI experts) and whether other procedural aspects of a trial complicate the more basic model of lay judgment discovered and presented here.

CONCLUSIONS

This study provides the first experimental evidence about physicians' potential liability for using AI in precision medicine. We find that two factors reduce lay judgment of liability: following standard care, and following the recommendation of AI tools. These results provide guidance to physicians who seek to reduce liability, as well as a response to recent concerns that the risk of liability in tort law may slow the use of AI in precision medicine. Contrary to the predictions of those legal theories, the experiments suggest that the view of the jury pool is surprisingly favorable to the use of AI in precision medicine.

ACKNOWLEDGEMENTS

We thank Camille Glauss, Patrick Krebs, and Reto Seiler for excellent legal research assistance; Atif Ghulam Nabi and Justin MacPherson for excellent programming assistance; and Dr. Henry Kim and Dr. Ben Nyblade for their advice on statistical methods. For comments on the paper, we thank Glenn Cohen, Allison Hoffman, and Stephen Latham, as well as seminar audiences at ETH Zurich and Yale University.

REFERENCES

[1] Price WN, Gerke S, Cohen IG. Potential liability for physicians using artificial intelligence. *JAMA*. 2019; 322:1765-1766.

[2] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med*. 2019; 25:44-56.

[3] Geistfeld M. Does tort law stifle innovative medical treatments? *Jotwell* 2015; https://torts.jotwell.com/does-tort-law-stifle-innovative-medical-treatments/. Accessed on September 19, 2020.

[4] Peters P. The quiet demise of deference to custom: malpractice at the millennium. *Washington & Lee Law Review* 2000; 57:163-205.

[5] Thone v. Reg'l W. Med. Ctr., 275 Neb. 238, 245, 745 N.W.2d 898, 905 (2008).

[6] Vidmar M. Juries and medical malpractice claims: Empirical facts versus myths. *Clin. Orthop. Relat. Res.* 2009; 467:367-375.

[7] Froomkin AM, Kerr I., Pineau J. When AIs outperform physicians: Confronting the challenges of a tort-induced over-reliance on machine learning. *Arizona Law Review* 2019; 61:33-99.

[8] Yuan Y, MacKinnon DP. Robust Mediation Analysis Based on Median Regression, *Psychological Methods.* 2014; 19(1):1-20.

[2] Rucker DD, Preacher KJ, Tormala ZL, Petty RE. Mediation Analysis in Social Psychology: Current Practices and New Recommendations. *Social and Personality Psychology Compass*. 2011:359-371.

[9] Dietvorst, BJ, Simmons JP, Massey C. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 2015; 144:114-126.

[11] Tobia K. How people judge what is reasonable. *Alabama Law Review* 2018; 70: 293-359.

Disclosure

There are no conflicts of interest.

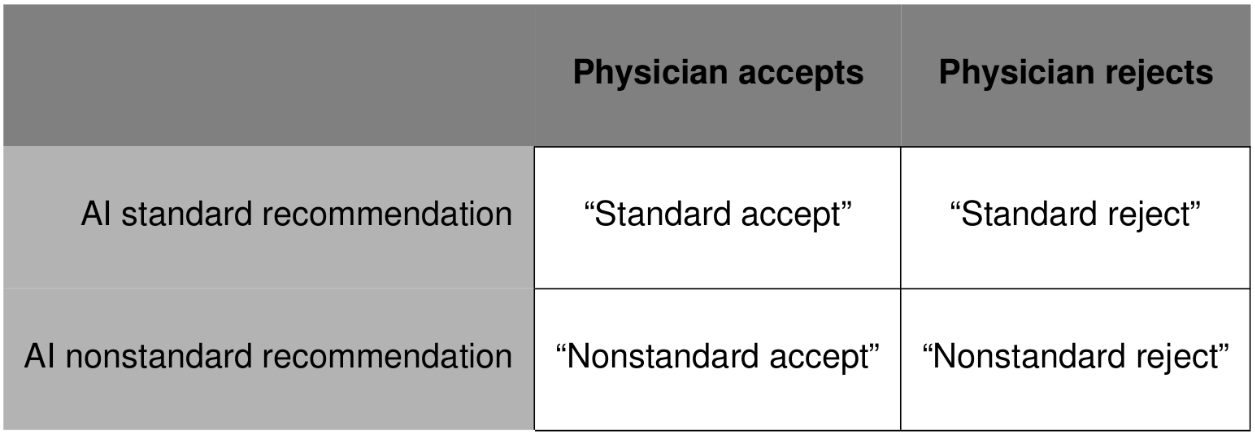
Key Points

Question: How can physicians minimize liability risk when using AI systems?

Pertinent Findings: A representative experimental study of 2,000 U.S. adults finds that two factors affect liability assessments: (1) whether standard, rather than nonstandard, care was provided; and (2) whether the AI advice was accepted.

Implications for Patient Care: Accepting AI advice reduces physician liability, offering a legal benefit even when there is a nonstandard care AI recommendation (all else equal).

Figures



	Physician accepts	Physician rejects
AI standard recommendation	"Standard accept"	"Standard reject"
AI nonstandard recommendation	"Nonstandard accept"	"Nonstandard reject"

Figure 1. Experimental Design, Crossing Recommendation (Standard, Nonstandard) with Decision (Accept, Reject)

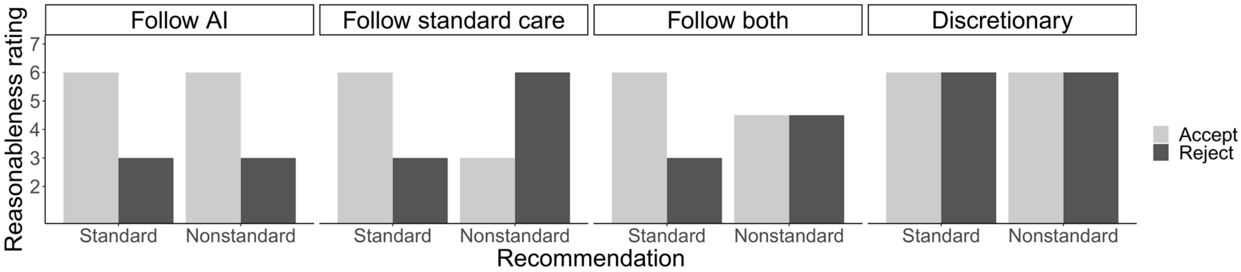
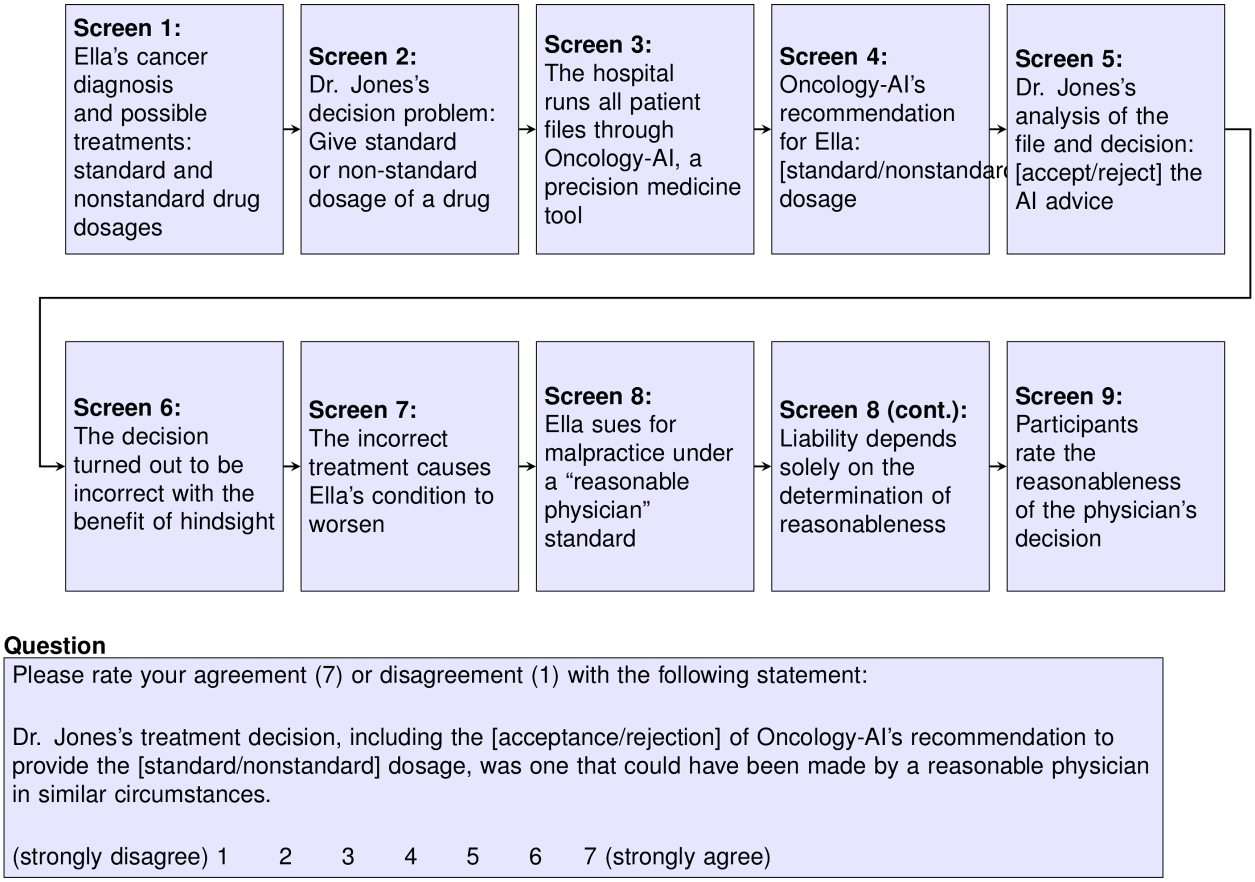


Figure 2. Experimental Predictions of Four Models


Screen 1: Ella's cancer diagnosis and possible treatments: standard and nonstandard drug dosages
Screen 2: Dr. Jones's decision problem: Give standard or non-standard dosage of a drug
Screen 3: The hospital runs all patient files through Oncology-AI, a precision medicine tool
Screen 4: Oncology-AI's recommendation for Ella: [standard/nonstandard] dosage
Screen 5: Dr. Jones's analysis of the file and decision: [accept/reject] the AI advice
Screen 6: The decision turned out to be incorrect with the benefit of hindsight
Screen 7: The incorrect treatment causes Ella's condition to worsen
Screen 8: Ella sues for malpractice under a "reasonable physician" standard
Screen 8 (cont.): Liability depends solely on the determination of reasonableness
Screen 9: Participants rate the reasonableness of the physician's decision
Question

Please rate your agreement (7) or disagreement (1) with the following statement:

Dr. Jones's treatment decision, including the [acceptance/rejection] of Oncology-AI's recommendation to provide the [standard/nonstandard] dosage, was one that could have been made by a reasonable physician in similar circumstances.

(strongly disagree) 1 2 3 4 5 6 7 (strongly agree)

Figure 3. Schematic flow chart of vignette

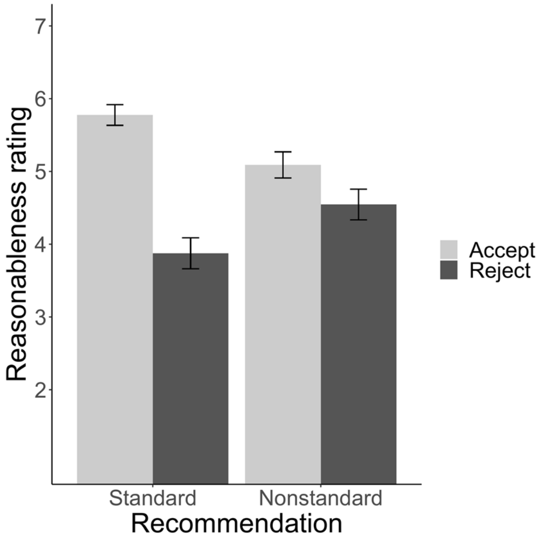


Figure 4. Mean ratings of Reasonableness, by condition. Error bars indicate 95% confidence intervals.

Table 1: Pairwise t-tests

				Standard Reject	Nonstandard Accept	Nonstandard Reject
Scenario	N	Mean	SD	t	t	t
Standard Accept	401	5.77	1.46	15.03*** d = 1.14 [.97, 1.30]	5.89*** d = .43 [.28, .57]	9.80*** d = .76 [.60, .92]
Standard Reject	311	3.87	1.91		-8.60*** d = -.67 [-.83, -.51]	-4.39*** d = -.36 [-.52,-.20]
Nonstandard Accept	360	5.09	1.74			3.86** d = .31 [.15, .46]
Nonstandard Reject	284	4.55	1.81			

** indicates p < .001, *** indicates p < .0001, with 95% Confidence Intervals; the two post-hoc comparisons (Standard Reject to Nonstandard Accept and Reject) reflect Bonferonni-corrected *p* values for two post-hoc tests).

Appendix

Table of Contents

I. Vignettes

Below is the text displayed across multiple screens to the participants in the four conditions. Variation is indicated by color-coded highlighting to indicate the Advice treatment [Standard/Nonstandard] and the Physician Decision [Accept/Reject]. In two cases the text shown will depend on the interaction of the two factors, so it is highlighted with gray, but the reader can infer which would be displayed given the combination of the two factors and the fact that *ex post* the physician made the wrong decision. For example, if the physician accepts the standard treatment advice and this was the wrong decision, we can infer that the AI system's advice was wrong and that the nonstandard treatment would have been the correct treatment option.

For screenshots and complete individual examples for each of the 2x2 treatment conditions, see https://osf.io/zyejh/?view_only=b0158bcde6a64d0da3a977b44b0610ca.

Screen 1

Ella has been diagnosed with ovarian cancer. The recommended medical treatment is to administer a chemotherapeutic drug, bevacizumab. There are two possible doses of the drug that could be administered. The standard care is the treatment that works best for most patients, but less well for others. The nonstandard care is the treatment that works best for some patients, but less well for most others.

Screen 2

• The standard care, which is the care that would be best for most patients, is to administer 15 milligrams of the drug per kilogram of body weight, every three weeks. Given that Ella weighs 60 kilograms, that recommendation translates into 900 milligrams every three weeks.

• The nonstandard care, which is the care that would be best for a small number of patients, is to administer a higher dosage: 75 milligrams per kilogram. That would translate into 4500 milligrams every three weeks for Ella.

Dr. Jones is choosing a treatment for Ella. The decision concerning whether to offer the standard care or the nonstandard care is difficult.

Screen 3

The hospital where Dr. Jones works runs all patient files through an AI treatment tool called Oncology-AI. Oncology-AI has all of the relevant medical approvals and is skilled in analyzing patients' data to estimate whether standard care or nonstandard care would likely be more successful.

Screen 4

Ella's data has been run through Oncology-AI, and the results have been included in Ella's file. Oncology-AI recommends that, for Ella, [standard care/nonstandard care] is most appropriate.

Screen 5

Based on his experience and careful analysis of the patient's file, Dr. Jones decides to provide standard care. That is, he [accepts/rejects] the recommendation of Oncology-AI. Dr. Jones provides Ella with [the standard treatment of 900 milligrams/the nonstandard treatment of 4500 milligrams] every three weeks.

Screen 6

It turns out that Dr. Jones's decision to follow the recommendation was the wrong choice for Ella. The recommendation of Oncology-AI was [incorrect/correct]. Rather than [the standard treatment of 900 milligrams/the nonstandard treatment of 4500 milligrams] every three weeks, Ella should have been given [the nonstandard treatment of 4500 milligrams/the standard treatment of 900 milligrams] every three weeks.

Screen 7

The incorrect treatment choice causes Ella's condition to worsen.

Screen 8

Now imagine that Ella has brought a lawsuit against Dr. Jones for medical malpractice.

Both Ella and Dr. Jones have agreed that Dr. Jones's treatment choice (which turned out to be incorrect) caused Ella's condition to worsen and that she was, in fact, harmed by that treatment choice.

In the state in which Dr. Jones and Ella reside, the key remaining question that determines whether Dr. Jones is liable in malpractice for the injury is whether "*a reasonable physician*" in similar circumstances could have made the same treatment decision as Dr. Jones.

Screen 9

Please rate your agreement (7) or disagreement (1) with the following statement:

Dr. Jones's treatment decision, including the acceptance of Oncology-AI's recommendation to provide the standard dosage, was one that could have been made by a reasonable physician in similar circumstances.

1 (strongly disagree) … 7 (strongly agree)

Screen 10

Please rate your agreement (7) or disagreement (1) with the following statement:

Dr. Jones's treatment decision, including the acceptance of Oncology-AI's recommendation to provide the standard dosage, was one that could have been made by a reasonable physician in similar circumstances.

1 (strongly disagree) … 7 (strongly agree)

II. Exclusions

1367 participants correctly answered both comprehension check questions correctly and were included in the main analysis, as outlined in the study pre-registration. An additional 11 data points contained duplicate IDs, inconsistent IDs, or completed the study in under 30 seconds. As such, in the exclusionary analyses, those 11 participants were also excluded. We also exclude two participants who took the survey twice, likely due to a technical error. The primary analyses are conducted using these stringent exclusion criteria. However, we also conducted our primary pre-registered analyses without applying exclusion criteria and the results are robust (see https://osf.io/zyejh/?view_only=b0158bcde6a64d0da3a977b44b0610ca).

III. ANOVA results, controlling for age, race, and gender

The main effect of Decision and Recommendation * Decision interaction are both robust when controlling for participants' self-reported age, race, and gender. Race included seven categories. Gender included four: male, female, non-binary, and prefer not to respond.

Table A1: ANOVA Table

	Model 1	Model 2
Recommendation	$F(1,\ 1352) = 0.00,\ \eta p^2$=.00	$F(1,\ 1335) = 0.05,\ \eta p^2$=.00
Decision	$F(1,\ 1352) = 167.71,\ \eta p^2$=.11**	$F(1,\ 1335) = 172.02,\ \eta p^2$=.11**
Recommendation * Decision	$F(1,\ 1352) = 51.68,\ \eta p^2$=.04**	$F(1,\ 1335) = 49.73,\ \eta p^2$=.04**
Age		$F(1,\ 1335) = 4.57,\ \eta p^2$=.00
Gender		$F(3,\ 1335) = 1.71,\ \eta p^2$=.00
Race		$F(6,\ 1335) = 1.55,\ \eta p^2$=.01

* indicates $p < .05$, ** indicates $p < .001$. Model 2 includes Bonferonni-corrected p-values for the three additional comparisons.

IV. Additional exploratory analyses for “ideal” and “average” physician measures

After evaluating reasonableness, each participant was presented with exploratory questions about ideal and average physicians: “Do you think *an ideal physician* in similar circumstances could have made the same treatment decision as Dr. Jones, including to [accept/reject] the [standard/nonstandard] recommendation?” and “Do you think *most physicians* in similar circumstances could have made the same treatment decision as Dr. Jones, including to [accept/reject] the [standard/nonstandard] recommendation?”

Table A2 reports paired t-tests, indicating that the three measures diverged significantly in the standard-reject and nonstandard-reject conditions, but not in the standard-accept and nonstandard-accept conditions.

Table A2: Paired t-tests

Standard Accept

Variable	N	Mean	SD	Ideal	Reasonable
Average	401	5.77	1.31	t=.58, d=.03 (-.06, .13)	t=.93, d=.05 (-.05, .14)
Ideal	401	5.80	1.35		t=-.44, d=-.02 (-.11,.08)
Reasonable	401	5.77	1.46		

Standard Reject

Variable	N	Mean	SD	Ideal	Reasonable
Average	311	3.39	1.80	t=-3.32,** d=-.19 (-.30,-.08)	t=-4.77,*** d=-.27 (-.38,-.16)
Ideal	311	3.66	1.92		t=2.15,*d=.12 (.01,.23)
Reasonable	311	3.87	1.91		

Nonstandard Accept

Variable	N	Mean	SD	Ideal	Reasonable
Average	360	4.95	1.68	t=.33, d=.02 (-.08,.12)	t=-1.58, d=-.19 (-.19,-.02)
Ideal	360	4.93	1.71		t=-1.78, d=-.09 (-.20,.01)
Reasonable	360	5.09	1.74		

Nonstandard Reject

Variable	N	Mean	SD	Ideal	Reasonable
Average	284	3.95	1.75	t=-2.18,* d=-.13 (-.25,-.01)	t=-5.30,*** d=-.31 (-.43,-.20)
Ideal	284	4.14	1.71		t=-3.50,** d=-.21 (-.33,-.09)
Reasonable	284	4.55	1.81		

* indicates $p < .05$, ** indicates $p < .01$, *** indicates $p < .001$, parentheses indicate 95% CIs

Figures A3 and A4 present results from a GLM mediation analysis in which the average and ideal measures are entered as multiple parallel mediators of the significant treatment effects of “Decision” (accept or reject the AI recommendation) and “Providing Standard” care (providing standard care by accepting standard advice or rejecting nonstandard advice; or providing nonstandard care by accepting nonstandard advice or rejecting standard advice). The analysis indicates that each of these two significant effects is partly mediated by both the average and ideal measures.

The GLM mediation analysis was conducted in Jamovi version 1.2., with the jammGLM command, which is built on the lavaan command for R. The analysis used 95% confidence intervals computed with the bootstrap percentiles method (1,000 bootstraps). Although computationally intensive, bootstrapping methods are more general and generate more reliable estimates than a standard mediation analyses.[1]

As Figure A3 indicates, both effects (the main effect of Decision; and "Provide Standard," the Decision * Recommendation interaction) are partially mediated by the average and ideal measures. We note that these results should be interpreted with caution, as "proving" mediation requires measuring all mediators and suppressors without error. Given the difficulty of measuring variables perfectly, we note that the direct versus total effect comparisons should be interpreted cautiously.[2]

Figure A3. Mediation Path Model

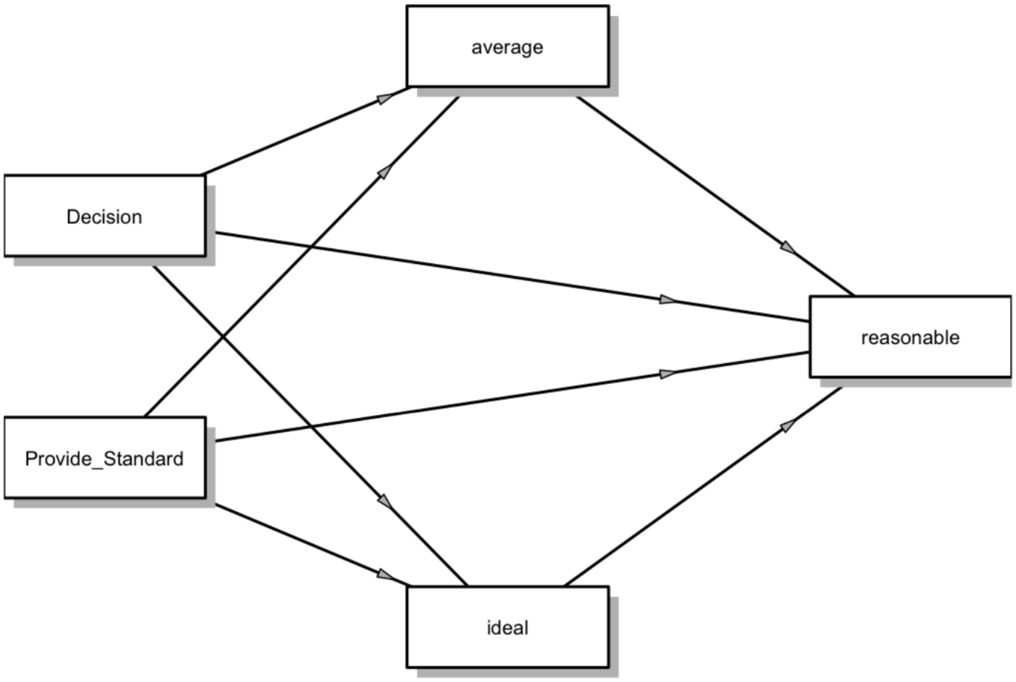
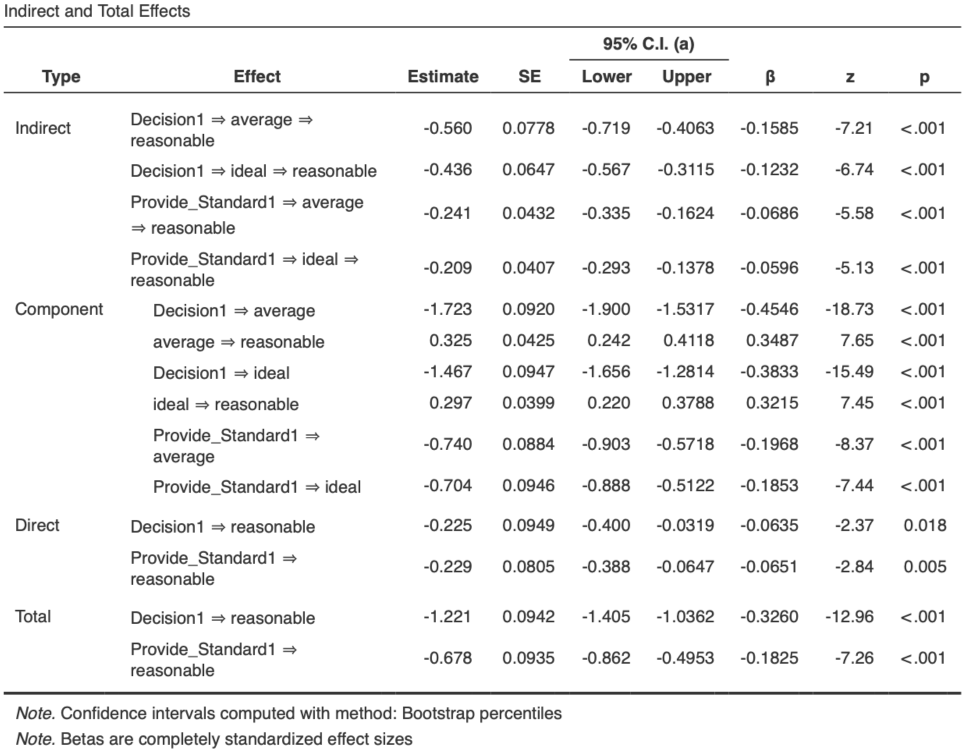


Figure A4. Indirect and Total Effects of Mediation Analysis



Indirect and Total Effects

				95% C.I. (a)				
Type	Effect	Estimate	SE	Lower	Upper	β	z	p
Indirect	Decision1 ⇒ average ⇒ reasonable	-0.560	0.0778	-0.719	-0.4063	-0.1585	-7.21	<.001
	Decision1 ⇒ ideal ⇒ reasonable	-0.436	0.0647	-0.567	-0.3115	-0.1232	-6.74	<.001
	Provide_Standard1 ⇒ average ⇒ reasonable	-0.241	0.0432	-0.335	-0.1624	-0.0686	-5.58	<.001
	Provide_Standard1 ⇒ ideal ⇒ reasonable	-0.209	0.0407	-0.293	-0.1378	-0.0596	-5.13	<.001
Component	Decision1 ⇒ average	-1.723	0.0920	-1.900	-1.5317	-0.4546	-18.73	<.001
	average ⇒ reasonable	0.325	0.0425	0.242	0.4118	0.3487	7.65	<.001
	Decision1 ⇒ ideal	-1.467	0.0947	-1.656	-1.2814	-0.3833	-15.49	<.001
	ideal ⇒ reasonable	0.297	0.0399	0.220	0.3788	0.3215	7.45	<.001
	Provide_Standard1 ⇒ average	-0.740	0.0884	-0.903	-0.5718	-0.1968	-8.37	<.001
	Provide_Standard1 ⇒ ideal	-0.704	0.0946	-0.888	-0.5122	-0.1853	-7.44	<.001
Direct	Decision1 ⇒ reasonable	-0.225	0.0949	-0.400	-0.0319	-0.0635	-2.37	0.018
	Provide_Standard1 ⇒ reasonable	-0.229	0.0805	-0.388	-0.0647	-0.0651	-2.84	0.005
Total	Decision1 ⇒ reasonable	-1.221	0.0942	-1.405	-1.0362	-0.3260	-12.96	<.001
	Provide_Standard1 ⇒ reasonable	-0.678	0.0935	-0.862	-0.4953	-0.1825	-7.26	<.001

Note. Confidence intervals computed with method: Bootstrap percentiles

Note. Betas are completely standardized effect sizes

These exploratory analyses provide some insight into the future of the law concerning AI in medicine. Recent work in cognitive science indicates that lay judgment of what is reasonable is driven by both what people think is common and what people think is good. Our exploratory findings are consistent with that research. If this is right, we would predict that as AI-use becomes more common among physicians, jurors will see AI-use as more reasonable.

V. References

[1] Yuan Y, MacKinnon DP. Robust Mediation Analysis Based on Median Regression, *Psychological Methods.* 2014; 19(1):1-20.

[2] Rucker DD, Preacher KJ, Tormala ZL, Petty RE. Mediation Analysis in Social Psychology: Current Practices and New Recommendations. *Social and Personality Psychology Compass*. 2011; 5/6:359-371.