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

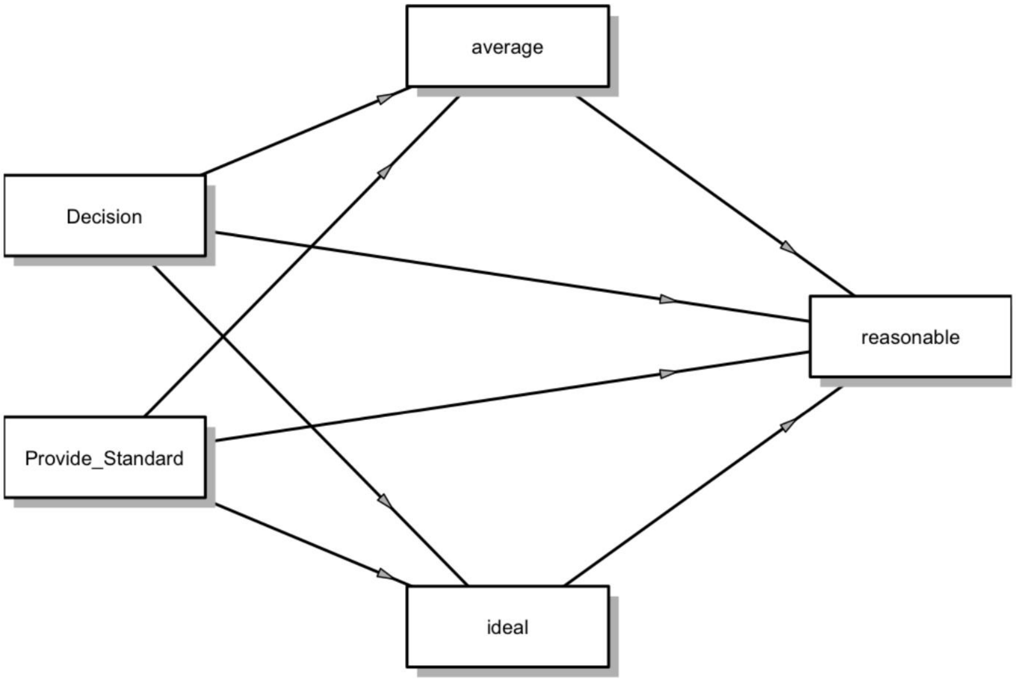
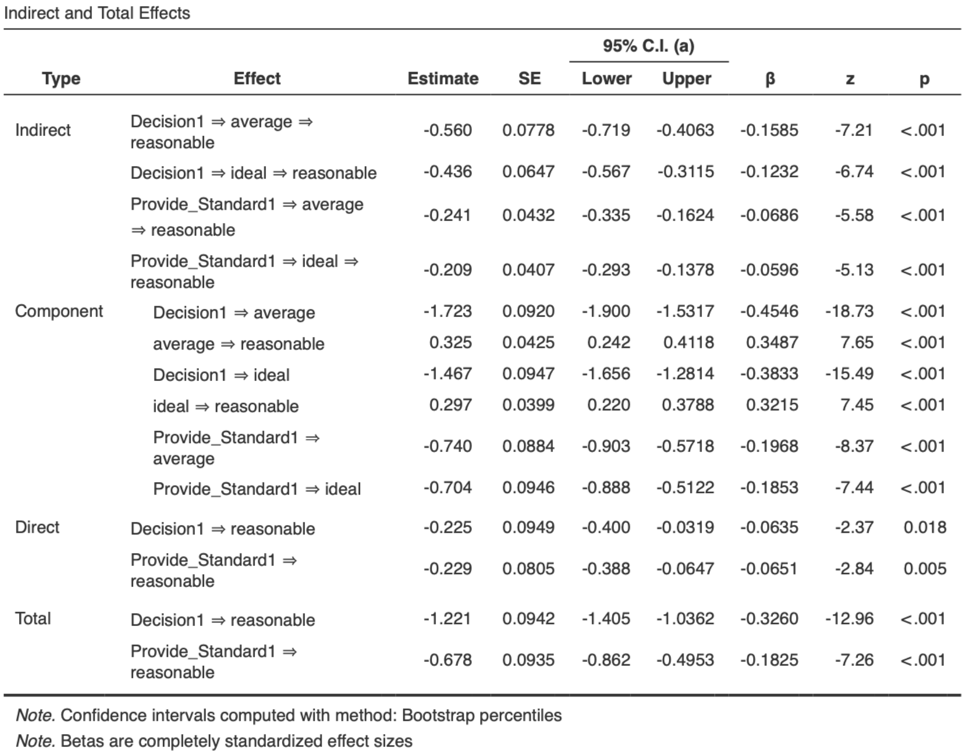


Figure A4. Indirect and Total Effects of Mediation Analysis



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.