Appendix A: Randomization of incorrect AI responses across periods

The sequence of task presentation was randomized for each participant. This randomization allowed for 36! (i.e., more than 3.7×10^{41}) possible (raw) sequences. However, we used the constraint of not having two incorrect AI response occur in direct sequence. For each participant, new permutations were generated until this condition was satisfied. Figure A1 provides an overview of a randomized sequence of tasks alongside the associated *Willingness to rely* stated by the participant.

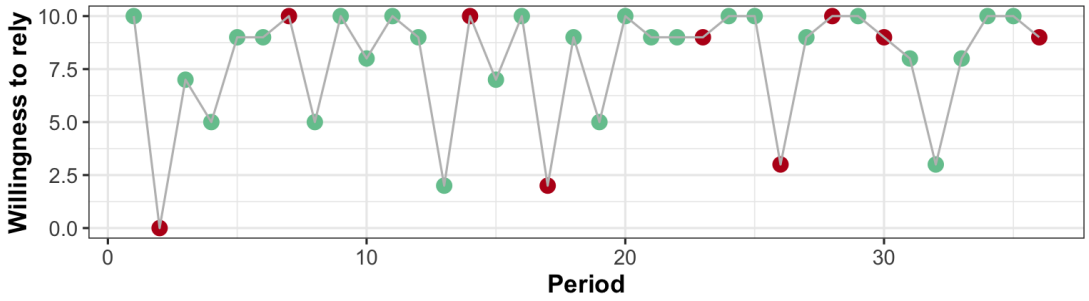

Figure A1: Example of *willingness to rely* stated for a random sequence of **correct** and **incorrect** task responses for one participant across all 36 periods

To assess the *balance* of correct and incorrect AI responses across periods, we consider the share of incorrect responses per period compared to expectation (Figure A2).

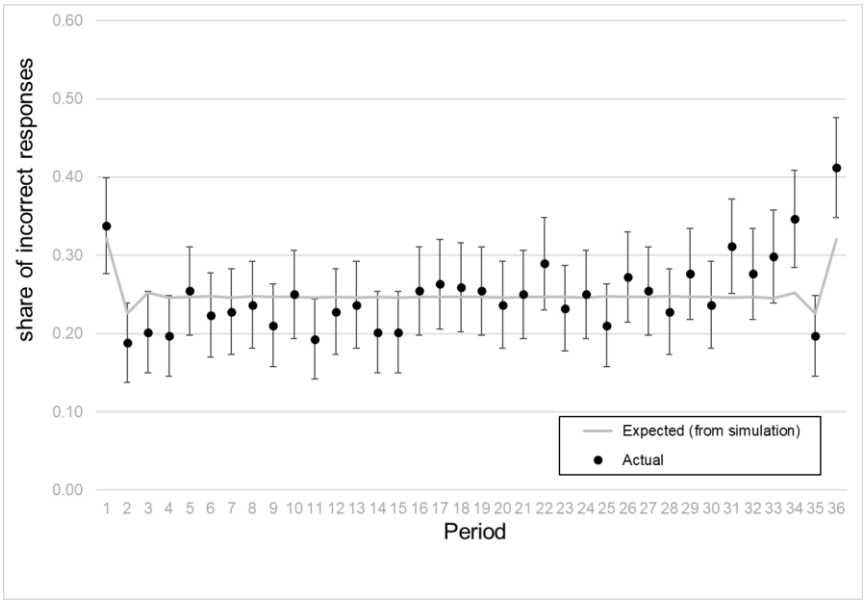

Figure A2: Expected and observed share of incorrect responses per period (error bars indicate 95% confidence intervals)

As can be seen there, these shares are balanced fairly well across periods around the 25% mark. We calculated the expected shares of incorrect responses per period based on a simulation of 10,000,000 random sequences of which around 7.3% satisfied the non-consecutiveness constraint. Note that this constraint had a slight effect on expected shares in the first and last periods. These values turn out to be somewhat higher as an incorrect response in those periods only “blocks” one other period whereas in any other period, two periods are blocked.

Appendix B: Category and task breakdown

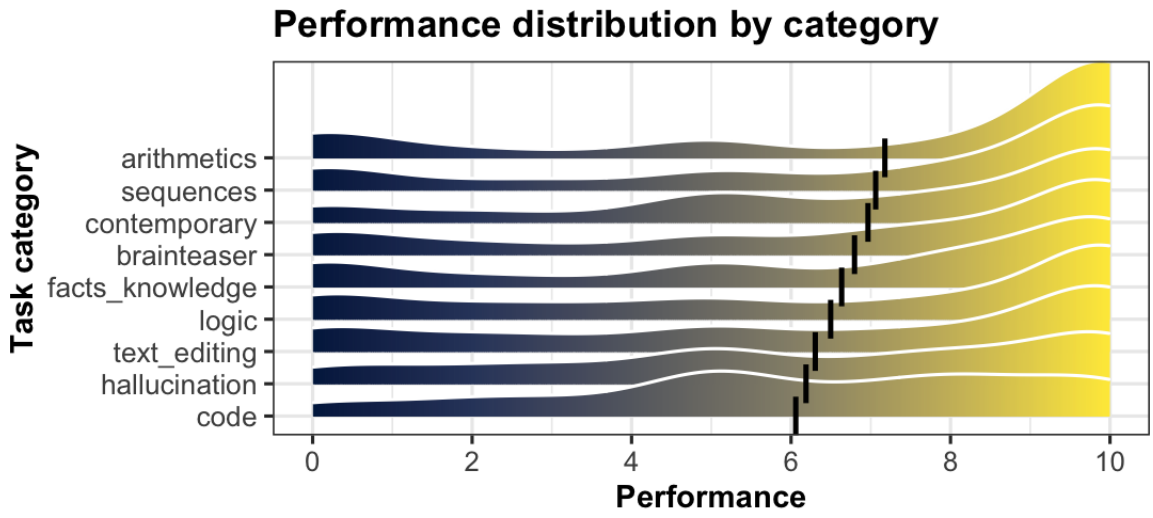

Figure B1: Distribution *of performance* per category across participants (vertical bars indicate category mean)

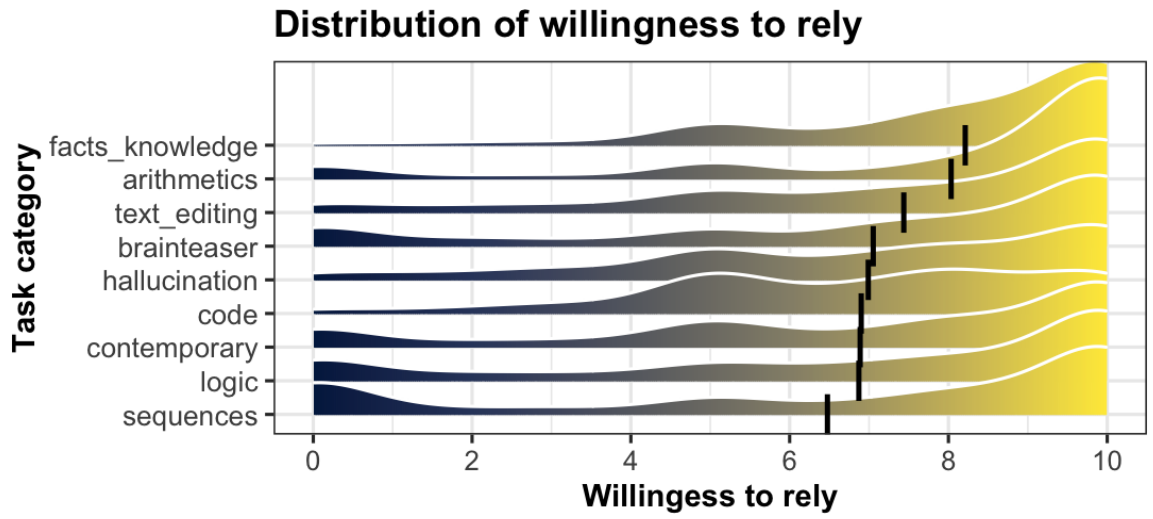

Figure B2: Distribution of *willingness to rely* per category across participants (vertical bars indicate category mean)

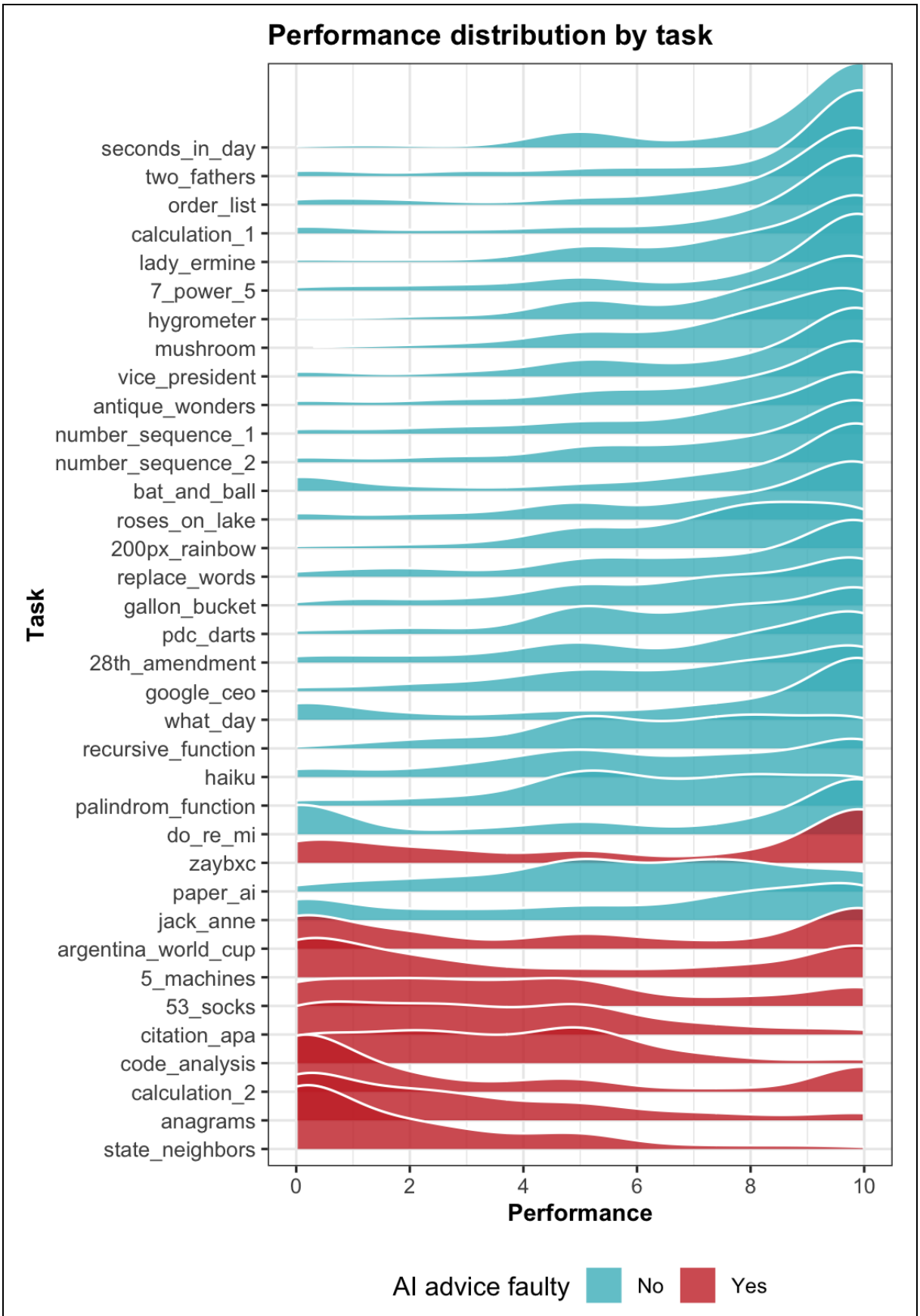

Figure B3: Distribution of *performance* per task (ordered by mean *performance* across participants)

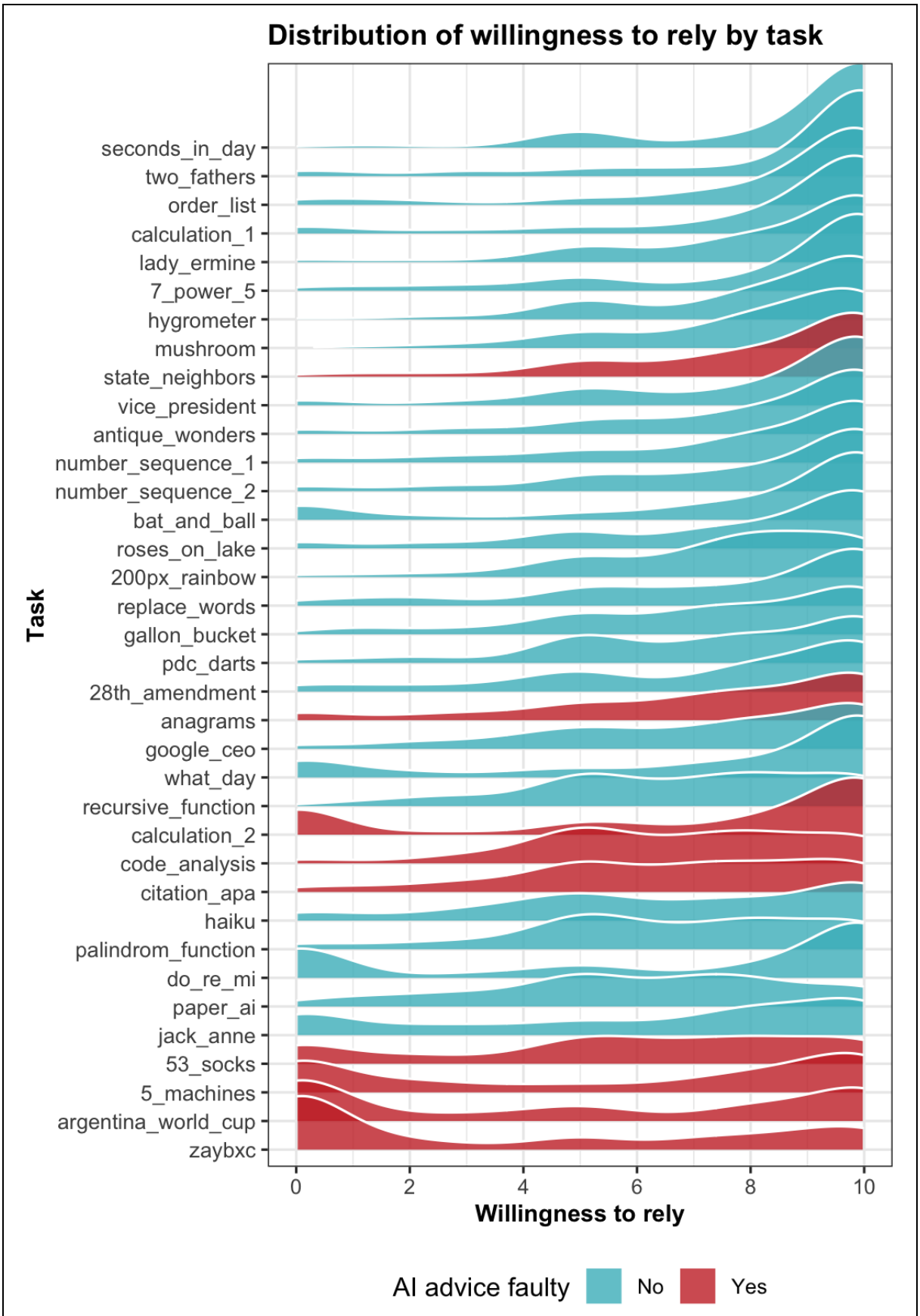

Figure B4: Distribution of *willingness to rely* by task (ordered by mean *willingness to rely* across participants)

Appendix C: Measurement instrument

Table C1: Items and scales in pre- and post-experiment survey

Section	Item	Wording	Response options
Pre-survey	AI skepticism (pre)	I am skeptical about text-generative AI (such as ChatGPT).	0 - strongly disagree 1 - disagree 2 - somewhat disagree 3 - neutral 4 - somewhat agree 5 - agree 6 - strongly agree
	AI curiosity (pre)	I am curious about text-generative AI.	
	AI trust (pre)	I trust text-generative AI.	
	AI risk (pre)	I believe text-generative AI to be a risk.	
	AI innovation (pre)	I believe text-generative AI to be a great innovation.	
Post-survey	AI skepticism (post)	I am skeptical about text-generative AI (such as ChatGPT).	0 - strongly disagree 1 - disagree 2 - somewhat disagree 3 - neutral 4 - somewhat agree 5 - agree 6 - strongly agree
	AI curiosity (post)	I am curious about text-generative AI.	
	AI trust (post)	I trust text-generative AI.	
	AI risk (post)	I believe text-generative AI to be a risk.	
	AI innovation (post)	I believe text-generative AI to be a great innovation.	
	ChatGPT familiarity	What statement best describes your experience with ChatGPT before participating in this study?	0 - This is the first time I have heard about ChatGPT 1 - I have heard about ChatGPT but don't really know what it is/does 2 - I know what ChatGPT is but have never used it myself before 3 - I have used ChatGPT at least once before 4 - I use ChatGPT every now and then (i.e., a couple of times per month) 5 - I use ChatGPT very frequently (i.e., a couple of times per week)
	ChatGPT usage time	If you have used ChatGPT before, how long have you been using it?	0 - Less than 1 week 1 - More than 1 week, less than 1 month 2 - More than 1 month, less than 2 months 3 - More than 2 months, less than 4 months 4 - More than 4 months
	ChatGPT account	Do you have a registered ChatGPT Account?	0 - No 1 - Yes
	ChatGPT paid account	Do you have a subscription to ChatGPT?	
	Risk attitude	Are you generally a person who is fully prepared to take risks or do you try to avoid taking risks?	0 - not at all willing to take risks - *up to* - 10 - very willing to take risks
	Trust in technology I	I usually trust machines until there is a reason not to.	0 - strongly disagree 1 - disagree 2 - somewhat disagree 3 - neutral 4 - somewhat agree 5 - agree 6 - strongly agree
	Trust in technology II	In general, I would rely on a machine to assist me.	
	Trust in technology III	My tendency to trust machines is high.	
	Technology proficiency	I consider myself to be proficient in understanding and using technology.	
	Computer knowledge	I feel very knowledgeable about computers and software.	
	AI knowledge	I feel very knowledgeable about artificial intelligence.	
	Realism check AI performance	I think that the provided prompts and responses were a good representation of text-generative AI performance.	
	Realism check AI job usage	The scenario presented here (i.e., reading and assessing AI text responses) seems plausible in general and may be a common part of future knowledge-work.	
	Manipulation check time pressure	I felt a lot of time-pressure to evaluate the AI's responses.	
	Manipulation check briefing	I felt I was briefed well about the limitations and weaknesses of text-generative AI"	
	performance self assessment (absolute)	I believe I correctly identified whether the AI responses were correct or wrong with great accuracy.	
	performance self assessment (relative)	I believe I performed better than most other participants.	

Appendix D: Manipulation and realism checks

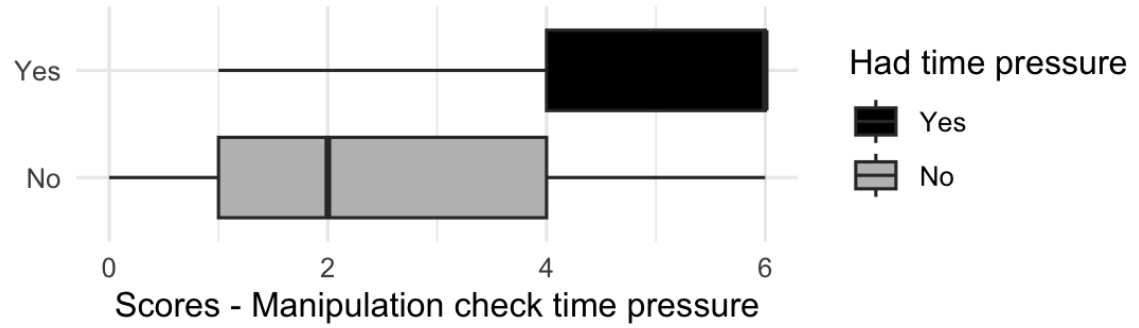

Figure D1: Manipulation Check (Time Pressure)

Question: *"I felt a lot of time-pressure to evaluate the AI's responses"*

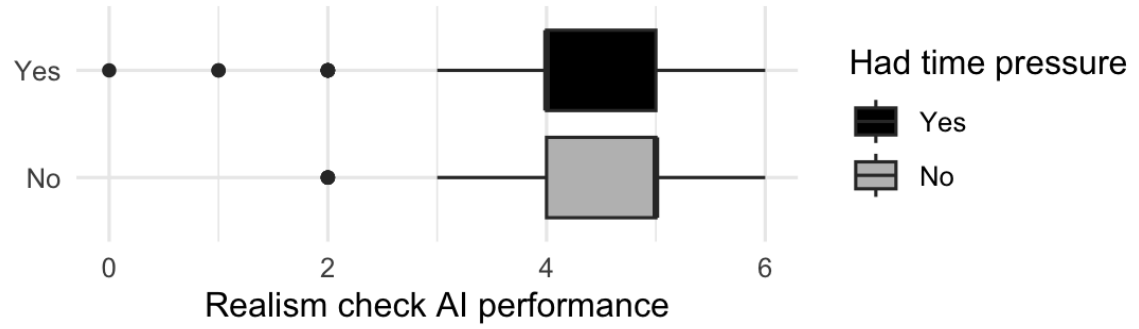

Figure D2: Realism check 1 (AI performance)

Question: "*I think that the provided prompts and responses were a good representation of text-generative AI performance"*

Figure D3: Realism check 2 (AI job usage)

Question: *"The scenario presented here (i.e., reading and assessing AI text responses) seems plausible in general and may be a common part of future knowledge-work"*

Appendix E: Statistics for relaxed definition of *choice univocity*

Table E1: Relaxed definition of *choice univocity*

Choice univocity metric	**Degree of relaxation**	**Overall univocal choices**	Breakdown of univocal choices per treatment group		Overall breakdown by "direction"	
			Control group (n=4,068)	**Treatment group** (n=4,140)	**Univocal willingness to rely**	**Univocal willingness <u>not</u> to rely**
strict	± 0 (i.e., stated willingness to rely 0 or 10)	*43.21%*	*50.05%*	*36.25%*	*7.18%*	*36.04%*
		3,547	2,072	1,475	589	2,958
medium	± 1 (i.e., stated willingness to rely 0,1, 9 or 10)	*57.14%*	*63.20%*	*51.18%*	*9.42%*	*47.72%*
		4,690	2,571	2,119	773	3,917
relaxed	± 2 (i.e., stated willingness to rely 0, 1, 2, 8, 9 or 10)	*71.36%*	*75.81%*	*66.98%*	*11.88%*	*59.48%*
		5,857	3,084	2,773	975	4,882

Note: A total of 8,208 responses were given by 228 respondents across 36 tasks each, thereof 4,140 (115 respondents) in the treatment group (time pressure) and 4,068 (113 respondents) in the control group.

Appendix F: Sensitivity / Specificity

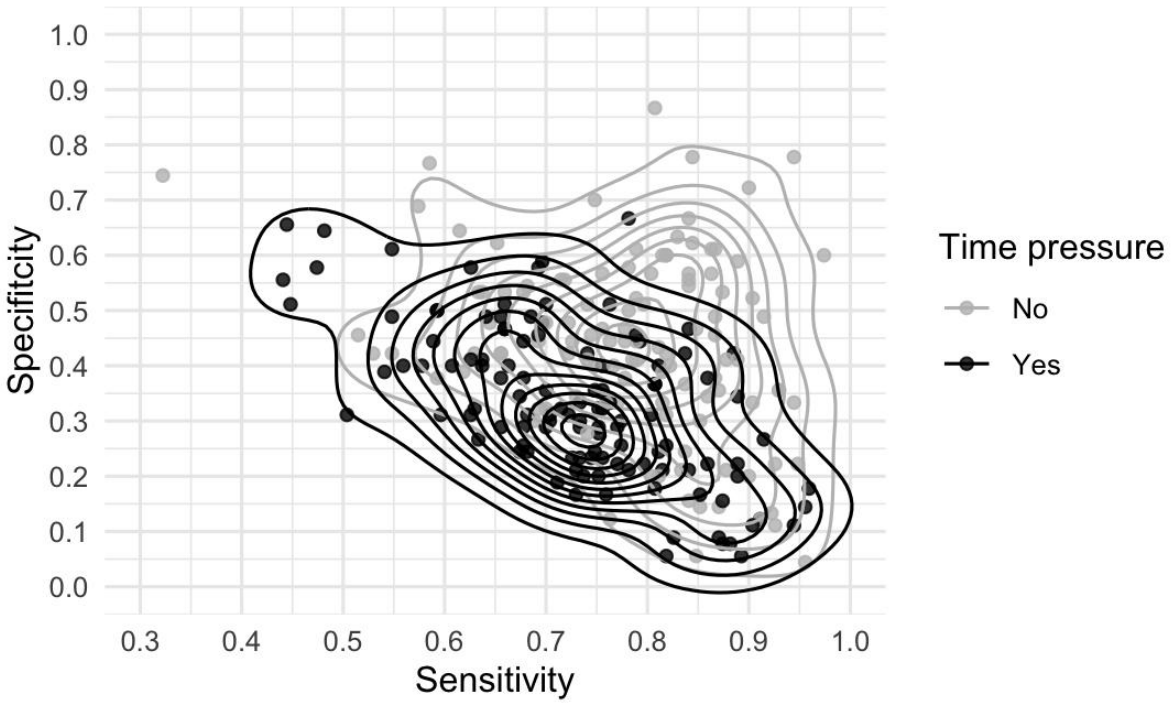

Figure F1*: Sensitivity* (i.e., proportion of correct AI responses accurately identified as correct by humans) and *specificity* (i.e., the proportion of incorrect AI responses accurately identified as incorrect by humans) for both treatment groups