Should My Agent Lie for Me? A Study on Attitudes Towards Deceptive AI

Keywords: Deceptive AI; deception; value-alignment; user study; future-of-work

Extended Abstract

Purpose: Is it acceptable for machines to be dishonest? On the one hand, there is vast research in social sciences suggesting that deception has a detrimental effect on trust and cooperation [Schweitzer et al., 2006]. Given this, there is a strong argument that we must be prepared to hold deceptive machines (that replace humans in day-to-day social contexts) accountable [Dignum, 2019]. On the other hand, other research shows that a prosocial form of deception can increase trust, when the intentions are good [Levine and Schweitzer, 2015]. And other behavioural research shows that machines can be better at inducing cooperation, if they use some form of dishonesty [Ishowo-Oloko et al., 2019]. Perhaps the answer to the above question depends on the situation, and approaching this with concrete examples may suggest that the answer is contextual. For example, should an AI-powered medic deceive its patient about a life-threatening diagnosis if this would maximise the patient's mental well-being for the rest of their remaining lifetime? What about an AI-powered personal secretary facing the dilemma of lying about their employer's daily schedule? Or situations where a machine does not provide any false information, but it creates misleading advertisements that exploits the public's poor maths skills (e.g., thinking that 1/3 is less than 1/4 [Tversky and Kahneman, 1983])? Does it really matter if the machine is taking a role that is perceived to have deception as an essential part of it (e.g., lawyers) [Hangstler, 1993]? Assuming that there are situations where machines designed with the ability to deceive are seen in a positive light or are even desirable, the question is then: where do people draw the line between acceptable and unacceptable dishonesty by machines?

We believe that human perception is an important part of the puzzle, since it would affect people's adoption and trust in these machines. To answer the above question, and to understand human perception of the potentially deceptive behaviour of machines in different social contexts, we started with a study. Our end goal is to explore different factors in combination and develop a computational model that captures the crucial factors (and their interactions) that have a meaningful influence on people's perception of what makes dishonest behaviour by machines acceptable.

We asked the following research questions: **Q1**: When would AI deception be perceived as more permissible by humans? **Q2**: Would humans trust AI agents capable of deception? **Q3**: Would humans want to adopt or buy AI agents capable to deceive? **Q4**: Who would humans hold responsible? And we hypothesised that agent types, deception roles, and demographic differences play a role in how people assign moral permissibility to deceptive agent behaviour, how much people trust deceptive agents, how willing people are to buy the services of deceptive agents, and how people assign responsibility to the entities involved in the deception.

Methods: The experiment used a 2 x 3 between-subject design: dishonest agent [human vs. AI] x Beneficiary and target of the lie [deception of someone for the subject's benefit vs deception of the subject for someone else's benefit vs deception of someone for someone else's

benefit. (control)]. 424 Participants based in the US, aged between 21 - 66 (M = 33, SD = 9), N_{female} = 183(43.16%), N_{male} = 241(56.84%) completed this story-based user study online. They were randomly assigned to one of the 6 conditions. In each condition, each subject read 5 stories representing 5 different applications and answer questions about moral permissibility, trust, willingness to buy, and responsibility.

Results: In line with our preregistration, a series of linear regression models (see Table 1) were fitted to predict the effects of agent, beneficiary, age, gender, socio-economic status (SES), education, income, religiosity and political view on participants' moral permissibility (deceiver vs user), trust in agent, willingness to buy, responsibility assignment to different parties(the beneficiary vs the deceiver vs the AI maker), respectively (See Table 1). Our results found no significant effect of agent type and beneficiary on the attribution of moral permissibility, trust and willingness to buy. However, participants' self-indicated SES and religiosity level predict their attributions of moral permissibility (see Figure 1), trust (see Figure 2) and willingness (see Figure 3). Moreover, responsibility assignment varies by agent type and participants' demographic groups including those defined by their education, income level, religiosity and political views (see Figures 4, 5 and 6). Namely, the beneficiary of the deception received less responsibility when using deceptive AI rather than a deceptive human. Finally, more-religious participants generally assigned more responsibility to everyone involved in the deception (the beneficiary, the deceiver and the AI designer) compared to less-religious ones.

Impact: We believe our work has important implications on the design of algorithms and machines, especially those that are prepared to take central roles in our society. It would also raise further questions. For example, how do we make sure deceptive machines are ethically aligned with us, while reaping the benefits they might offer? And can we use our understanding of human perception about deceptive machines in order to mitigate the negative effects of the evolution of deceptive AI in increasingly hybrid societies?

References

[Dignum, 2019] Dignum, V. (2019). *Responsible artificial intelligence: how to develop and use AI in a responsible way*. Springer Nature.

[Hangstler, 1993] Hangstler, G. (1993). Vox populi-the public perception of lawyers. *American Bar Association Journal*, pages 60–65.

[Ishowo-Oloko et al., 2019] Ishowo-Oloko, F., Bonnefon, J.-F., Soroye, Z., Crandall, J., Rahwan, I., and Rahwan, T. (2019). Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. *Nature Machine Intelligence*, pages 1–5.

[Levine and Schweitzer, 2015] Levine, E. E. and Schweitzer, M. E. (2015). Prosocial lies: When deception breeds trust. *Organizational Behavior and Human Decision Processes*, 126:88–106.

[Schweitzer et al., 2006] Schweitzer, M. E., Hershey, J. C., and Bradlow, E. T. (2006). Promises and lies: Restoring violated trust. *Organizational behavior and human decision processes*, 101(1):1–19.

[Tversky and Kahneman, 1983] Tversky, A. and Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review*, 90(4):293.

Figures and Tables

Table 1: Overview of Regression Results

	Dependent variable						
	moral_deceiver[1]	moral_user[2]	trust[3]	will[4]	resp_benef[5]	resp_dec[6]	resp_AImaker[7]
agentAI	0.02 (0.18)	0.04 (0.21)	-0.01 (0.21)	0.03 (0.21)	-0.53** (0.19)	-0.30 (0.18)	
beneficiarysomeone	-0.27 (0.22)	-0.33 (0.25)	-0.35 (0.25)	-0.28 (0.25)	-0.42 (0.23)	-0.15 (0.21)	-0.14 (0.32)
beneficiaryyou	0.09 (0.23)	0.07 (0.26)	0.06 (0.27)	-0.04 (0.26)	-0.16 (0.24)	-0.11 (0.23)	0.29 (0.34)
age	-0.01 (0.01)	0.002 (0.01)	-0.01 (0.01)	-0.003 (0.01)	-0.001 (0.01)	0.02 (0.01)	0.01 (0.01)
genderMale	0.14 (0.20)	-0.03 (0.23)	-0.08 (0.23)	0.10 (0.23)	-0.14 (0.21)	-0.16 (0.20)	-0.20 (0.30)
SES	0.27*** (0.06)	0.17* (0.07)	0.21** (0.07)	0.16* (0.07)	0.11 (0.06)	0.02 (0.06)	0.09 (0.09)
edu.HighSchool	0.41 (0.58)	-0.55 (0.65)	-0.22 (0.67)	-0.37 (0.66)	-1.44* (0.61)	-1.25* (0.57)	-0.74 (1.44)
edu.undergrad	0.10 (0.57)	-0.24 (0.64)	0.09 (0.65)	0.12 (0.65)	-0.67 (0.60)	-0.23 (0.55)	0.36 (1.43)
edu.postgrad	-0.11 (0.60)	-0.64 (0.68)	-0.48 (0.69)	-0.25 (0.69)	-1.52* (0.64)	-0.96 (0.59)	-0.06 (1.45)
incomelvl.medium	0.003 (0.22)	-0.36 (0.25)	-0.24 (0.26)	-0.26 (0.26)	-0.41 (0.24)	-0.68** (0.22)	-0.07 (0.33)
incomelvl.high	0.48 (0.33)	0.09 (0.38)	0.21 (0.39)	0.11 (0.38)	-0.10 (0.35)	-0.30 (0.33)	-0.25 (0.48)
Religiosity	0.21*** (0.05)	0.28*** (0.06)	0.37*** (0.06)	0.31*** (0.06)	0.37*** (0.05)	0.24*** (0.05)	0.38*** (0.08)
PoliticalView	0.04 (0.05)	0.04 (0.06)	-0.04 (0.06)	0.01 (0.06)	0.06 (0.06)	0.18*** (0.05)	0.10 (0.08)
Constant	3.83*** (0.67)	4.29*** (0.75)	4.02*** (0.77)	4.13*** (0.76)	5.41*** (0.70)	5.33*** (0.65)	3.25* (1.47)
Observations	424	424	424	424	424	424	219
R^2	0.27	0.22	0.26	0.22	0.32	0.29	0.38
Adjusted R^2	0.24	0.19	0.24	0.19	0.30	0.27	0.34
Residual Std. Err.	1.85 (df = 410)	2.08 (df = 410)	2.13 (df = 410)	2.11 (df = 410)	1.95 (df = 410)	1.81 (df = 410)	1.92 (df = 206)
F Stat.	11.55*** (df = 13; 410)	8.66*** (df = 13; 410)	11.11*** (df = 13; 410)	8.68*** (df = 13; 410)	15.12*** (df = 13; 410)	13.14*** (df = 13; 410)	10.33*** (df = 12; 206)

Note: * p<0.05; **p<0.01; ***p<0.001

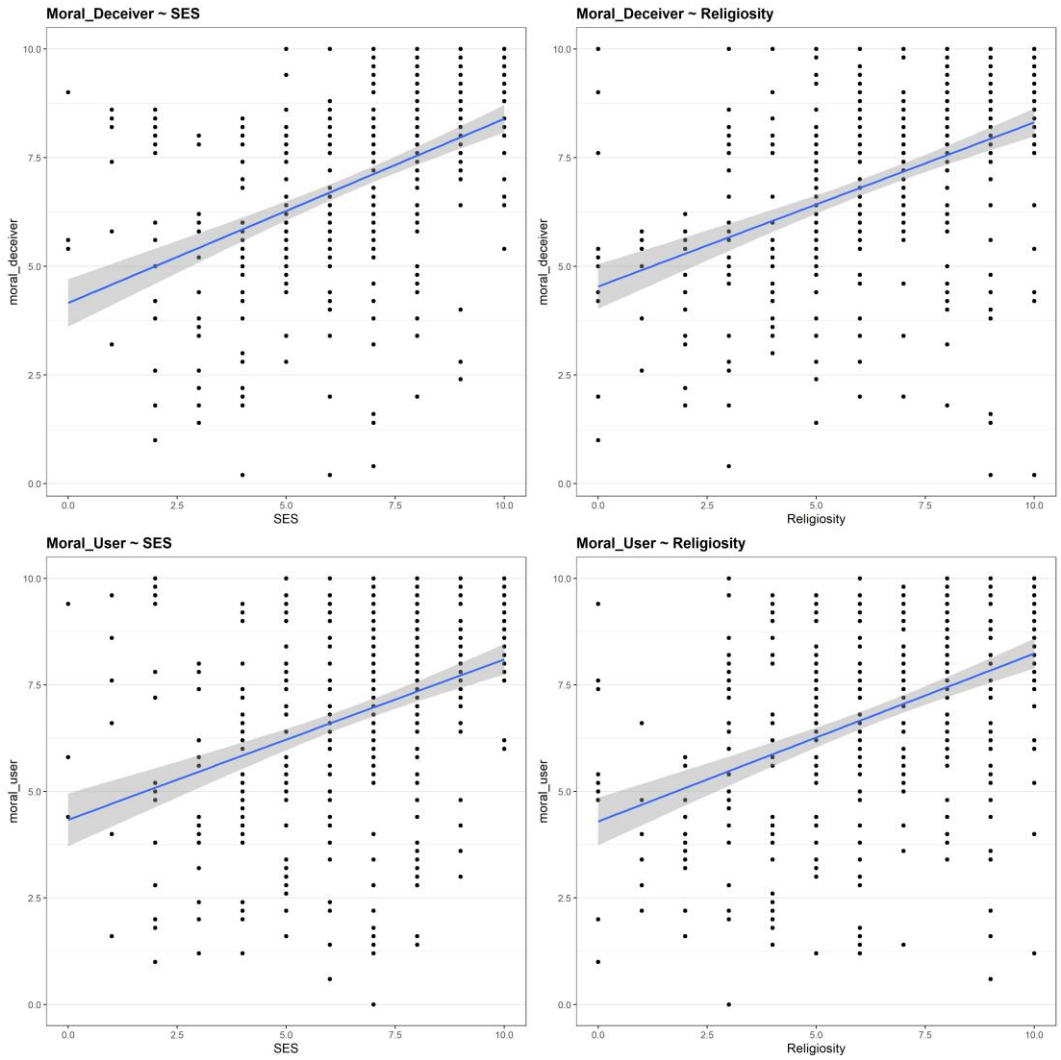

Figure 1: Moral Permissibility for Deceiver and User by SES and Religiosity. *The Predictive Associations of SES and Religiosity are positively related to Moral Permissibility for Deceiver and User.*

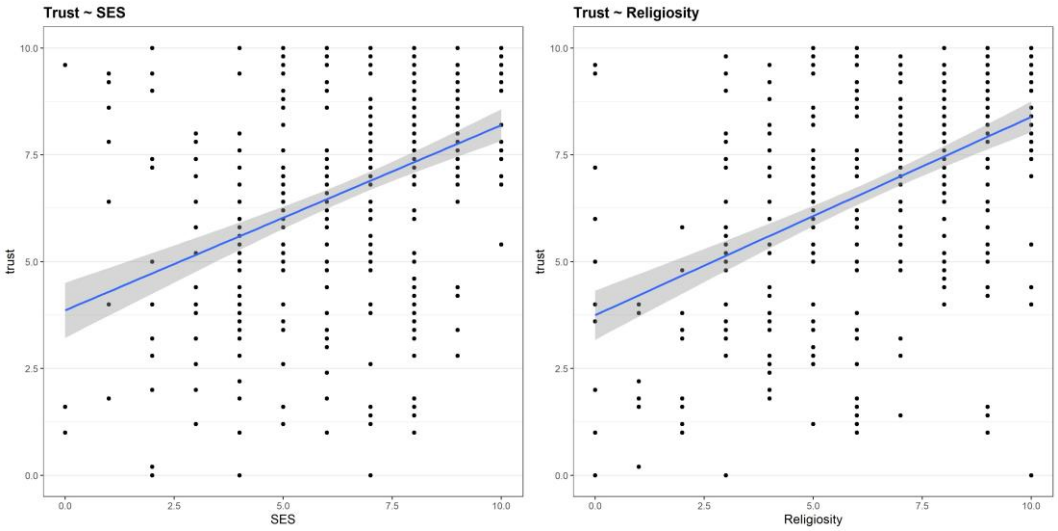

Figure 2: Trust in deceptive agents by SES and Religiosity. *The Predictive Associations of SES and Religiosity are positively related to trust.*

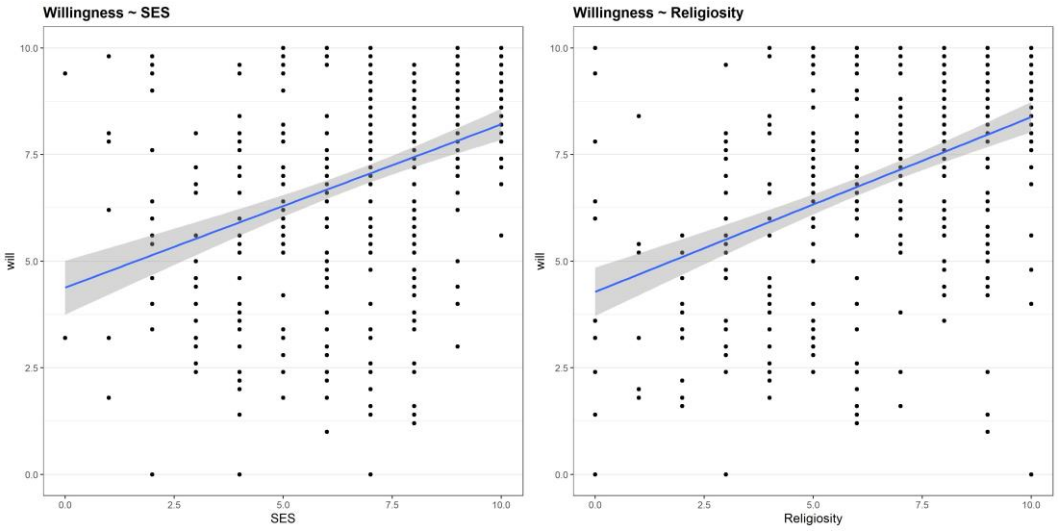

Figure 3: Willingness to Buy by SES and Religiosity. *The Predictive Associations of SES and Religiosity are positively related to willingness.*

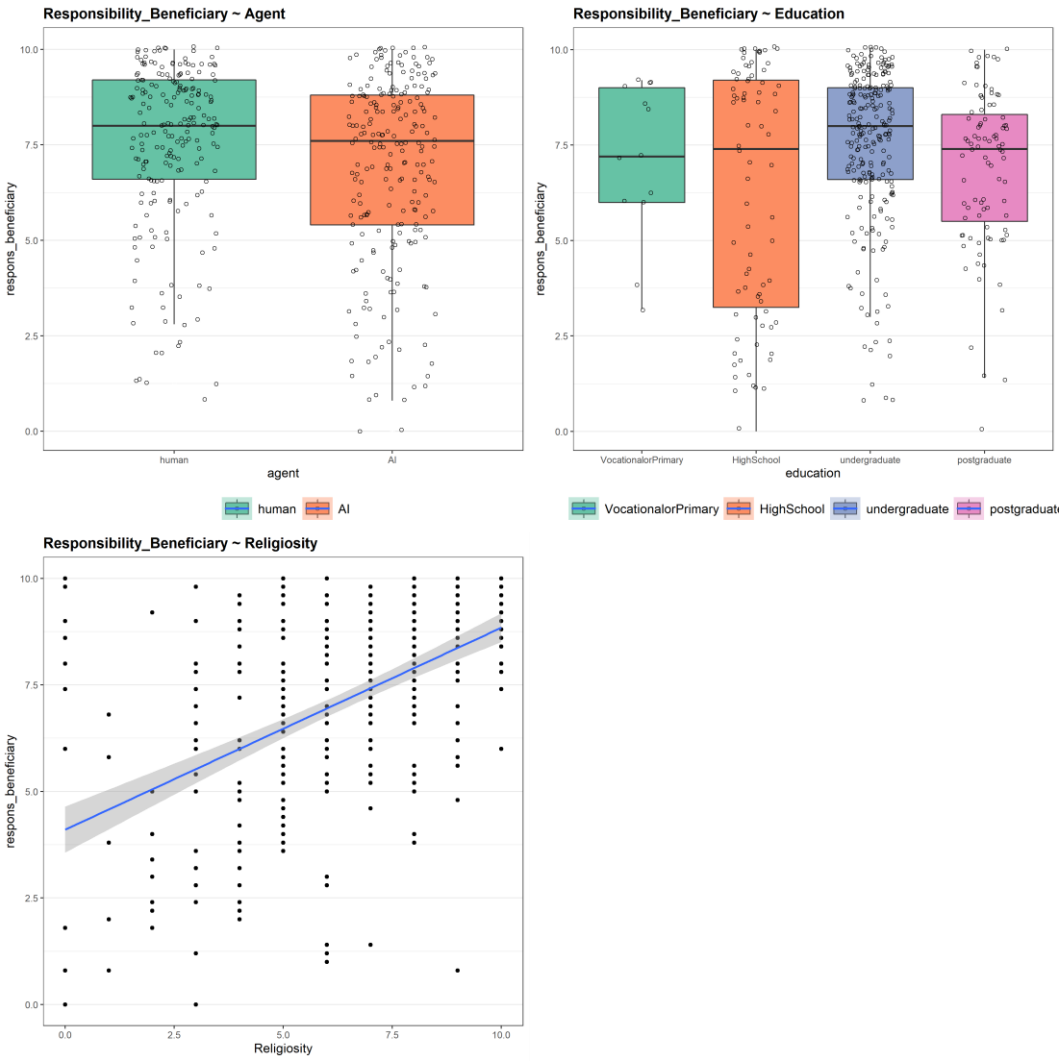

Figure 4: Responsibility for Beneficiary by Agent, Education and Religiosity. *Responsibility Assignment for Beneficiary is lower for deceptive AI, in participants have postgraduate degrees or with lower religiosity levels.*

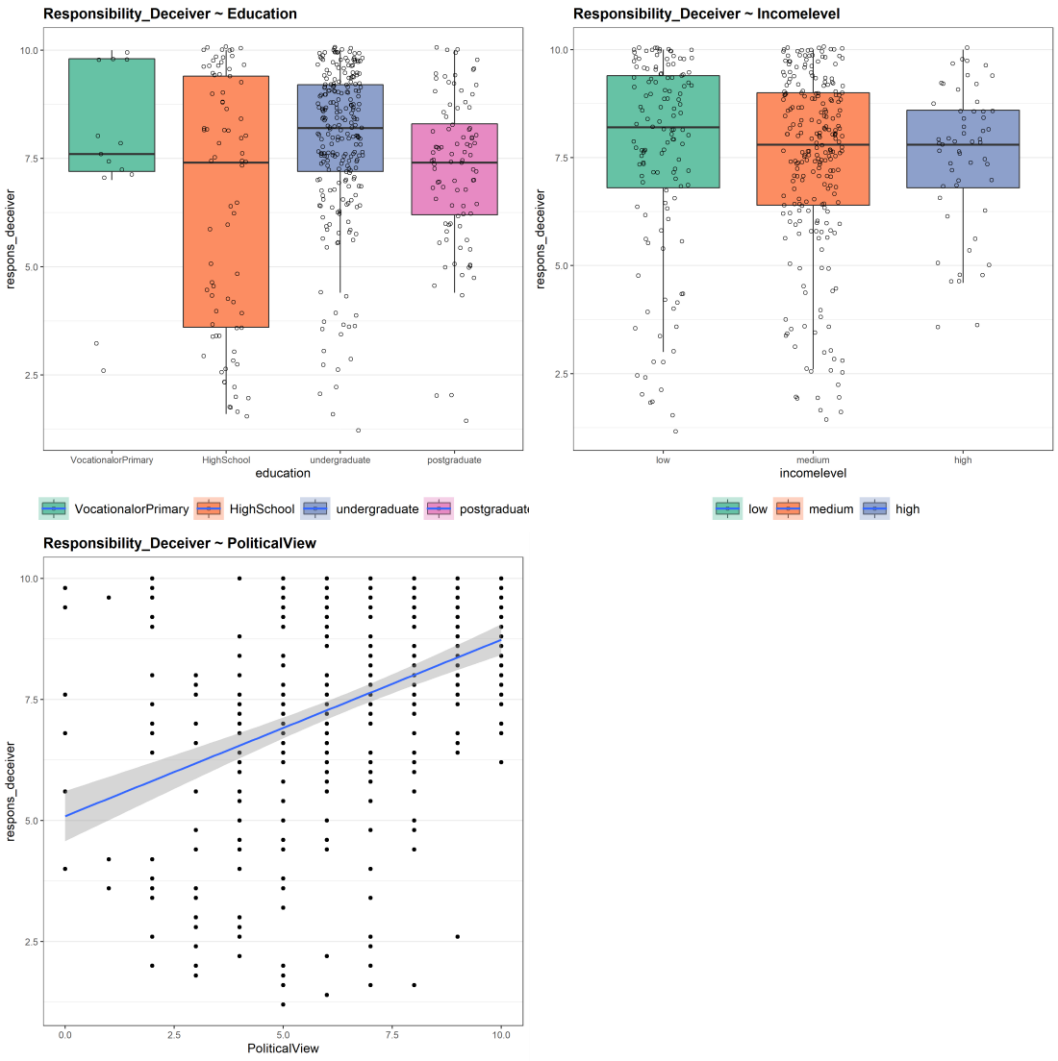

Figure 5: Responsibility for Deceiver by Education, Income Level, and Political View. *Responsibility Assignment for Deceiver is lower in participants have high school educations, medium income levels or more progressive political views.*

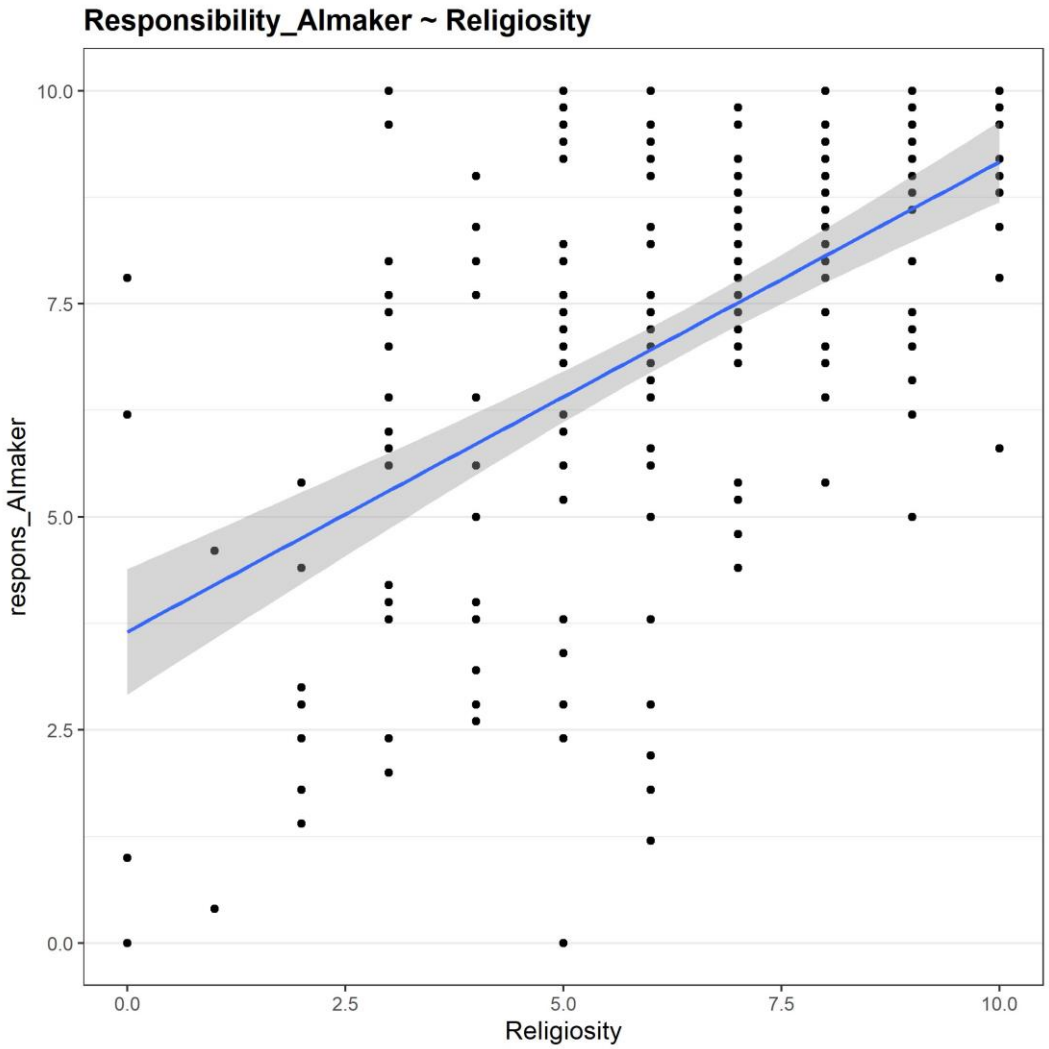

Figure 6: Responsibility for AI maker by Religiosity. *The Predictive Association of Religiosity is positively related to Responsibility Assignment for AI maker.*