Judging One's Own or Another Person's Responsibility in Interactions With Automation

Nir Douer and **Joachim Meyer**, Tel Aviv University, Israel

Objective: We explore users' and observers' subjective assessments of human and automation capabilities and human causal responsibility for outcomes.

Background: In intelligent systems and advanced automation, human responsibility for outcomes becomes equivocal, as do subjective perceptions of responsibility. In particular, actors who actively work with a system may perceive responsibility differently from observers.

Method: In a laboratory experiment with pairs of participants, one participant (the "actor") performed a decision task, aided by an automated system, and the other (the "observer") passively observed the actor. We compared the perceptions of responsibility between the two roles when interacting with two systems with different capabilities.

Results: Actors' behavior matched the theoretical predictions, and actors and observers assessed the system and human capabilities and the comparative human responsibility similarly. However, actors tended to relate adverse outcomes more to system characteristics than to their own limitations, whereas the observers insufficiently considered system capabilities when evaluating the actors' comparative responsibility.

Conclusion: When intelligent systems greatly exceed human capabilities, users may correctly feel they contribute little to system performance. They may interfere more than necessary, impairing the overall performance. Outside observers, such as managers, may overweigh users' contribution to outcomes, holding users responsible for adverse outcomes when they rightly trusted the system.

Application: Presenting users of intelligent systems and others with performance measures and the comparative human responsibility may help them calibrate subjective assessments of performance, reducing users' and outside observers' biases and attribution errors.

Keywords: human-automation interaction, decision making, warning systems, warning compliance

Address correspondence to Joachim Meyer, Tel Aviv University, Wolfson Building, Ramat Aviv, Tel Aviv, 69978, Israel; e-mail: jmeyer@tau.ac.il

HUMAN FACTORS
2022, Vol. 64(2) 359–371
DOI:10.1177/0018720820940516
Article reuse guidelines: sagepub.com/journals-permissions

INTRODUCTION

Artificial intelligence (AI) and advanced automation made it possible to create systems, in which computers and humans share the collection and evaluation of information, decision-making, and action implementation. In these systems, human responsibility has become equivocal.

Responsibility is a complex issue, involving role responsibility, causal responsibility, legal responsibility, and moral responsibility (Hart, 2008; Hart & Honor, 1985; Vincent, 2011). We focus on human *causal responsibility* when interacting with intelligent systems, which we define as the comparative human contribution to outcomes. For humans to have causal responsibility, they must be able to control the system and the resulting consequences (Noorman & Johnson, 2014).

As systems become more intelligent, there is a shift toward *shared control*, in which humans and computers jointly make decisions and control actions, or *supervisory control*, in which the human sets high-level goals, monitors the system, and only intervenes if necessary (Abbink et al., 2018). These types of control are used in complex sociotechnical systems, characterized by large problem spaces, highly coupled subsystems, advanced automation, interaction mediation via computers, uncertainty in the available data, and disturbance by unanticipated events (Vicente, 1999). AI algorithms increase these difficulties, as the internal processes can be opaque ("black box") and occasionally produce peculiar counterintuitive results (Castelvecchi, 2016; Scharre, 2016). Consequently, humans may no longer be able to control intelligent systems sufficiently to be considered fully responsible for the outcomes (Crootof, 2015; Cummings, 2006; Docherty et al., 2012; Sparrow, 2009), and the system (or its developers) may share some of the responsibility (Coeckelbergh, 2012; Johnson & Powers, 2005).

This leads to a "responsibility gap" in the ability to divide causal responsibility between humans and systems (Docherty et al., 2012; Johnson et al., 2014; Matthias, 2004).

We developed a Responsibility Quantification ("ResQu") model to compute measures of human causal responsibility in intelligent systems (Douer & Meyer, 2020a). Using information theory, we quantified human causal responsibility as the expected share of unique human contribution to the overall outcomes. The measure reflects characteristics of the operational environment, the system and the human, and the function allocation between them.

With binary classification systems, such as alerts or alarms, the model is reduced to relatively simple calculations. Let X denote the binary set of the human's possible actions, and Y denote the binary classification output from the system. Then, the ResQu model defines human responsibility as

$$Resp\left(X\right) \underset{=}{\text{def}} \frac{H(X/Y)}{H(X)} = \frac{H(X,Y)-H(X)}{H(X)} \quad (1)$$

where $H(X)$ is Shannon's entropy, which is a measure of uncertainty related to a discrete random variable X

$$H\left(X\right) = -\sum_{x\in\chi} p\left(x\right) log_2 p\left(x\right) \quad (2)$$

and $H(X/Y)$ is the conditional entropy, which is a measure of the remaining uncertainty about a variable X when a variable Y is known.

$$H\left(X/Y\right) = -\sum_{y\in\gamma} p\left(y\right) \sum_{x\in\chi} p\left(x/y\right) log_2 p\left(x/y\right) \quad (3)$$

$Resp(X) = 0$ if, and only if Y completely determines X, in which case, the human fully complies with the system's classifications. $Resp(X) = 1$ if, and only if, the human action selection X is independent of the system's classification result Y, in which case, the human is fully responsible for the output.

For computing the "theoretical responsibility," the ResQu model assumes that humans are perfectly rational and use the system optimally. This implies that the human will assume more responsibility when a system has inferior classification abilities and less responsibility with a system with superior abilities. However, people may interact nonoptimally with systems, and they may also misperceive their contribution to a process. We tested the predictive value of the model in controlled laboratory experiments (Douer & Meyer, 2020b). We demonstrated that the ResQu model is not only a theoretical model, but it can serve as a descriptive model for predicting human behavior ("measured responsibility") and the perception of one's own contribution ("subjective responsibility").

We applied the ResQu model to failure detections in a factory control room. We discovered that managers considered control room operators as more responsible for adverse outcomes than was justified (Douer et al., 2020). These subjective attributions resemble aspects of the "fundamental attribution error" in social psychology, in which observers tend to overestimate dispositional factors for behaviors of another actor, while underestimating situational and environmental explanations for these behaviors (Ross, 1977, 2018).

The attribution error arises, because the actor's behavior is the primary reference point for the observers, and situational constraints receive less attention (Andrews, 2001; Lassiter et al., 2002; Smith & Miller, 1979). Moreover, observers may have unrealistic expectations regarding actors' capabilities and behavior (Gilbert & Malone, 1995). People from individualistic (Western) cultures, who view themselves and others as independent agents, are more prone to make the error than people from collectivistic cultures, who are more influenced by contextual information (Choi et al., 1999; Markus & Kitayama, 1991).

When interacting with systems, users tend to blame the system for errors and adverse outcomes (Friedman, 1995; Madhavan & Wiegmann, 2007; Morgan, 1992). Differently from how they perceive their own errors, users see system errors as evidence for characteristics of the system, often ignoring temporary or uncontrollable exogenous factors (van Dongen & van Maanen, 2006). Managers, however, may wrongly attribute failures to employees' undesirable personality traits, ignoring the system contribution (Davison & Smothers, 2015; Douer et al., 2020; Rogoff et al., 2004; van Dyck et al., 2005). Hence, it is important for system designers and policymakers

to consider how different human and automation capabilities affect users' and observers' responsibility perceptions, as it may influence users' behavior and the way they are judged by others.

We report here an experiment on perceptions of causal responsibility by actors who used automated decision support systems and by observers who watched the actors' actions and the decision aid's performance. We designed the task so that participants could not perform the task well without aid. Either a more or a less-accurate decision aid supported participants in the task. Based on our literature review and our previous research, we pose several hypotheses:

Hypothesis 1 (H1): Actors will behave according to the ResQu model predictions, taking on significantly more responsibility with the less-accurate system than with the accurate one.

Hypothesis 2 (H2): Both actors and observers will realize that the accurate system has better capabilities than the less-accurate system.

Hypothesis 3 (H3): Actors' and observers' subjective responsibility perceptions will correspond to the actors' behavior. If H1 is true, these assessments will be significantly lower for the accurate system than for the less-accurate system.

Hypothesis 4 (H4): We predict attribution errors, especially for the interaction with the less-accurate system, when both the humans' and the system's capabilities are similarly poor, and many adverse outcomes are expected. Specifically, observers will overrate actors' causal responsibility for adverse outcomes, whileereas the actors will tend to assess systems with (similar) capabilities as significantly inferior to themselves.

THE EXPERIMENT

The experiment involved a binary decision, aided by a simple binary alert system, resembling decisions in industrial control rooms, flight decks, vehicles, medical systems, smart homes, and many other systems (Bregman, 2010; Cicirelli et al., 2016; Doi, 2007; Jalalian et al., 2013; Meiring & Myburgh, 2015). In this case, the ResQu model is reduced to relatively simple calculations and interpretations.

The goal of the binary classification is to determine which of two possible categories an item belongs to. One usually refers to rare events that need to be detected (malfunctions, a pathology, etc.) as the signal. We used a Gaussian Signal Detection Theory (SDT) model (Green & Swets, 1966) to define the probabilistic characteristics of the system and the human user. In terms of SDT, aided decision-making is the combined performance of the human and the system (Maltz & Meyer, 2001; Meyer, 2001; Sorkin & Woods, 1985; Sorkin, 1988). Both the human and the decision aid obtain different information, which is probabilistically related to the actual environmental state. This information allows some discrimination between the two states, but there is ambiguity left regarding the true state. The human and the system are imperfectly correlated because otherwise, they would be redundant.

The detection sensitivity in SDT is the detector's ability to distinguish between signal and noise. We will denote by d'_A and d'_H, respectively, the alert and human detection sensitivities (measured in standard deviations [SD] of the distributions of observed stimuli). The larger the detection sensitivity, the easier it is for the detector to distinguish between signals and noise. The response criterion defines the detector's tendency to classify events as signal or noise. The alert system has a preset response criterion, denoted by β_A, which is used to determine its binary output. When the human works alone, without the use of a decision aid, the optimal response criterion, β_H, that maximizes the expected payoffs is:

$$\beta_H{}^* = \frac{(1-P_s)}{P_s} \cdot \frac{(V_{CR}-V_{FA})}{(V_{Hit}-V_{Miss})} \qquad (4)$$

where P_S is the signal probability, $1 - P_S$ is the noise probability, V_{CR}, V_{FA}, V_{Hit}, and V_{Miss} represent, respectively, the payoffs for different response outcomes: Correct Rejection (CR; correctly responding "noise"), False Alarm (FA; falsely responding "signal"), Hit (correctly

responding "signal"), and Miss (falsely responding "noise").

The alert's output serves as additional input for the human, who uses it to judge the observed ambiguous stimulus with two different response criteria. With a reliable alert system, the human should adopt a lower cutoff point when an alarm is issued (i.e., increase the tendency to declare a signal) and a higher cutoff point when no alarm is issued (Robinson & Sorkin, 1985).

The human's differential adjustment of the cutoff points, according to the alert's output, can serve as a measure for the level of human trust in the system (Meyer, 2001; Meyer & Lee, 2013). If the human uses a single cutoff point, regardless of indications from the system, he or she obviously ignores the system's indications and has no trust in the system. The larger the difference between the cutoffs, the greater the trust and the weight given to the information from the system, and the lower share of unique human contribution (i.e., lower measured responsibility)

In the experiment, participants classified ambiguous stimuli, receiving the aid from one of two alert systems. One system's detection sensitivity exceeded that of the participants, whereas the other system's sensitivity was similar to that of the participants. We conducted the experiment on pairs of participants. In each pair, one participant (the "Actor") actively performed the aided classification task, whereas the other participant (the "Observer") observed passively.

For both alert systems, we computed the ResQu model's theoretical responsibility predictions for optimal behavior. We computed the actors' actual level of responsibility from their performance data and collected subjective responsibility assessments from both actors and observers.

Materials and Methods

The experiment was conducted in the "Interaction with Technology (IwiT) Lab" of the Industrial Engineering department at Tel Aviv University on groups of up to eight participants. The instructions described the experiment as a simplified simulation of a quality control task in a factory. A certain percentage of items the factory produces are defective. A quality control worker inspects and classifies each item and decides if it is "intact" or a "defect" that should be discarded. The worker decides, based on the height of a displayed rectangle, which was sampled from one of two overlapping distributions. Participants were told that the factory considers acquiring an alert system to support the classification task. The factory considers two candidate systems, which may differ in their classification accuracy.

The two participants in each pair were randomly assigned to the roles of the actor and observer. The actor sat at a computer. The observer sat behind and to the right of the actor and could see the stimulus, the alert indications, the actor's actions, and the outcomes. Each pair did the experiment with two alert systems, and participants were told that they will be asked to rate and compare the performance and contribution of the two candidate systems.

This research complied with the American Psychological Association Code of Ethics and was approved by the Ethics Committee for Research with Human Participants at Tel Aviv University. Informed consent was obtained from each participant.

Participants. Participants were 60 undergraduate students from the Tel Aviv University Faculty of Engineering (ages 20–29, median 23, 48% females). They were recruited through email invitations. Each participant received 40 Israeli New Shekels (ILS), about US$12, for taking part in the experiment. Conscientious performance was encouraged by the promise of an additional monetary award (100 ILS, about US$29) to a randomly selected participant, using the accumulated individual scores as weights.

Design and procedure. The experiment was conducted on desktop computers, with Intel® i7 3.4 GHz Processor, 8 GB RAM, NVIDIA® GeForce GT 610 Video Card, and 23-inch (56 cm) monitors. The experimental program was written in Python.

Figure 1 shows a schematic depiction of the experimental screen. It consisted of a 20 cm high and wide square at the center of the screen. Above the square were two fields, labeled

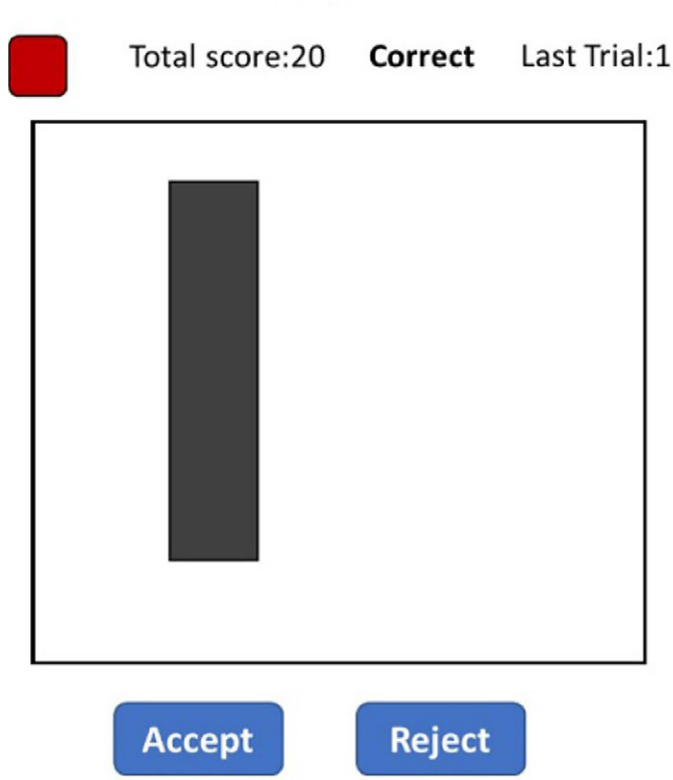

Figure 1. A schematic depiction of the experimental screen when there is an alert, the cumulative number of points is 20, and the participant chose a correct response in the last trial, which awarded an additional point.

"Total Score" and "Last Trial," which displayed the cumulative number of points and the number of points gained or lost in the last trial. Below the square were two action-selection buttons, labeled "Accept" (allowing an item) and "Reject" (discarding an item).

The ambiguous stimulus participants saw was a rectangle. It appeared inside the large square until the participant pressed one of the action-selection buttons. The rectangle had a fixed width. Its height was sampled either from a distribution of long or of short rectangles, representing, respectively, the length distributions of defective and intact items. In each trial, participants had to decide to which distribution the rectangle belonged. Similar to a method used in previous studies (Meyer, 2001), in each trial, the rectangle appeared at a different position inside the large square preventing participants to mark the cutoff point explicitly by, for instance, placing their finger on the screen. We controlled the human detection sensitivity by setting the overlap between the two distributions.

Indications from the aid were randomly determined according to preset probabilities for Hits and FA. Participants saw the indications from the aid in a small square at the top of the screen, which could be either red, indicating an alert, or green, indicating no alert. The indication from the aid appeared together with the rectangle, and it remained visible when the rectangle was shown.

Actors responded by clicking on either the "Accept" or the "Reject" button at the bottom of the screen, according to whether they thought that the rectangle belonged to the shorter or the

TABLE 1: Outcome Probabilities for the Two Alert Systems in Experiment 1

		Defect (Signal)		Intact (Noise)			
Type of Alert System	Parameters	Red (Hit)	Green (Miss)	Red (False Alarm)	Green (Correct Rejection)	PPV	NPV
Less-accurate	$d_A' = 1.0$, $\beta_A = 1$	69%	31%	31%	69%	60%	77%
Accurate	$d_A' = 2.3$, $\beta_A = 1$	87%	13%	13%	87%	82%	91%

Note. PPV = positive predictive value; NPV = negative predictive value.

longer distribution. After the response, the payoff for the trial appeared in the "Last Trial" field, and the "Score" field was updated. A feedback message, stating either "correct" or "incorrect," appeared for 2 seconds, and then the next trial began.

The experiment included two alert systems, one with the low detection sensitivity $d'_A = 1$ (the "less-accurate" system) and the other with the high detection sensitivity $d'_A = 2.3$ (the "accurate" system). In both cases, we assigned the participants a low detection sensitivity ($d'_H = 1$), by setting a matching overlap between the long and short distributions of the displayed rectangles.

Actors received 1 point for correctly classifying an intact or a defective item (which were 40% of the items). They lost 1 point for discarding an intact item and lost 2 points for not detecting a defective item. This payoff scheme reflects a factory's incentive not to deliver defective items to customers, which is stronger than the incentive not to discard intact items. In both alert systems, the response criterion was $\beta_A = 1$, equal to participants' optimal unaided response criterion, as calculated in Equation 4:

$$\beta_H^* = \frac{(1-P_s)}{P_s} \cdot \frac{(V_{CR} - V_{FA})}{(V_{Hit} - V_{Miss})} = \frac{0.6}{0.4} \cdot \frac{1-(-1)}{1-(-2)} = \frac{0.6}{0.4} \cdot \frac{2}{3} = 1 \quad (5)$$

Table 1 summarizes the outcome probabilities for the two systems and presents their positive and negative predictive values (PPV and NPV, respectively). These are the probabilities that an item was defective when the system indicated a defect (PPV) and that it was intact when the system indicated that it was intact (NPV).

For the above experimental settings, the ResQu model predicts, through Equation 1, optimal theoretical responsibility of 12% with the "accurate" alert system and 69% with the "less-accurate" alert system. Hence, the optimal prescribed actor behavior is to rely heavily on the accurate system's abilities and only moderately on the less-accurate system.

In two parts of the experiment, participants saw alerts from the "less-accurate" system and the "accurate" system. The order of examining the two alert systems was counterbalanced so that half of the pairs saw alerts from the accurate system first, and the other half saw the systems in reversed order. Actors performed 100 trials with each of the two alert systems, deciding on each trial whether to discard or allow the presented item. The 100 trials with each system were divided into two blocks, each with 50 trials. The participants were told that the first block of 50 trials was mainly for learning and gaining experience with their own and the system's abilities, and performance will be assessed according to the achievements in the second block.

After completing 100 trials with an alert system, the participants filled out a questionnaire, providing their subjective judgments on the actor's and system's detection capabilities and comparative human responsibility. In each question, the participants rated their level of

TABLE 2: Questions for Participants' Subjective Assessments

Factor	Question #	Question
Alert detection capabilities	Q1	The alert system could distinguish between intact and faulty items.
Human detection capabilities	Q2	The human could distinguish (without the aid of the alert) between intact and faulty items.
Human responsibility (human contribution to action selection)	Q3	The human used the indications from the alert system to select an action.
	Q4	When selecting actions, the human relied more on the indications from the alert system than on own detection abilities
	Q5	The alert system had a low contribution—the human could have similar performance without it
Concluding responsibility comparison	F1	On which alert system did the human rely more, when making decisions?

TABLE 3: Theoretical Predictions and Empirical Behavior of the Actors

Alert System	Outcomes as a Function of Alert Indications				Responsibility			SDT—Cutoffs Difference		
	Alert		No Alert							
d'_A	Hit	False Alarm	Hit	False Alarm	Theoretical Prediction	Measured	Diff.	Theoretical Prediction	Measured Mean	Diff.
1 (Less-accurate)	71%	50%	40%	16%	69%	81%	12%	1.6	1.1	−.5
2.3 (Accurate)	85%	62%	30%	10%	12%	49%	37%	3.9	1.8	−2.1

Note. SDT = Signal Detection Theory.

agreement on a scale between 1 (not at all) and 7 (very much). After using the second alert system, participants answered a concluding question. It asked for their impression on which of the two systems they had relied more, using a verbal scale that included seven options ("the first system by far," "the first system," "the first system slightly more," "no difference," "the second system slightly more," "the second system," "the second system by far").

Actors and observers completed identical questionnaires in different areas of the lab, so they could not discuss or influence each other's evaluations. They were also instructed not to communicate with each other during the experiment. Table 2 presents the questions and the factors to which they relate.

RESULTS

Table 3 shows the mean probabilities for actors' Hit and FA responses with and without an alert for the two systems, the computed levels of responsibility the actors assumed (the measured responsibility) and their cutoff difference, as well as the theoretical predictions for the responsibility and the cutoff difference. A two-way mixed analysis of variance, with the type of alert system as a within-subjects variable and the order in which the systems were examined as a between-subject variable. There was no significant main effect of the order or any significant interaction. Thus, we focus on system type. As predicted, actors assumed significantly more responsibility with the less-accurate

system, $F(1,28) = 55.10$, $p < .0001$; *Par.* $\eta^2 = .66$. The actors' measured responsibility was significantly closer to the theoretical prediction with the less-accurate system than with the accurate system, with which the actors assumed much higher-than-optimal responsibility, $t(29) = 6.07$, $p < .0001$.

In both systems, the cutoff difference was lower than optimal, implying that actors tended to under-trust the indications from the systems. Like the measured responsibility, the deviance was significantly larger with the accurate system, implying that with it, actors overestimated their own capabilities, causing them to select substantially nonoptimal cutoffs, $t(29) = 10.52$, $p < .0001$.

When the alert systems indicated a defect item, the actors' Hit and FA rates were significantly higher with the accurate system $t(29) = 4.77$, $p < .0001$ and $t(29) = 2.49$, $p = .01$, respectively. This shows that actors tended to comply more with the accurate system's alerts, whether they were correct or false. Without alert, the actors' Hit and FA rates were lower with the accurate system, $t(29) = 1.71$, $p = .10$ and $t(29) = 2.47$, $p = .01$, respectively, which implies higher Miss and CR rates. Hence, similarly, actors tended to comply more with the accurate system when there was no alert.

Questions Q3–Q5 measured participants' subjective assessments of the actors' responsibility. After reverse-scoring questions Q3 and Q4, we computed the reliability to measure the consistency of the questions. The analysis showed high reliability, with Cronbach's $\alpha = .84$, so we used the average of the questions as an estimate for subjective responsibility.

We analyzed the subjective assessments with a three-way mixed repeated measures analyses of variance (ANOVA), with the role of the participant ("Observer"/"Actor") and the alert type ("Accurate"/"Less-accurate") as within-subjects variables and the alert order as a between-subject variable. We excluded two outliers (one actor and one observer) because their mean score was more than two SDs from the mean. Table 4 summarizes the ANOVA results. Table 5 presents the mean values for the significant variables.

In the analysis of question Q1, the difference between the two alert system capabilities was significant with large effect size and so was the order of experiencing the systems. Also significant, but with smaller effect sizes, were interactions between the participant's role and the other two variables. Both types of participants rated the accurate alert significantly higher than the less-accurate system, especially when it was examined second. In most cases, both types of participants gave the less-accurate system similar low ratings, but when it was examined first, observers rated it significantly higher.

In the analysis of question Q2, both actors and observers evaluated the human detection capability as moderate, regardless of the alert type and the order. When working with

TABLE 4: Analysis of Variance Results for Questions Q1, Q2, Q3–Q5

	Q1: Alert Detection Capability		Q2: Human Detection Capability		Q3–Q5: Human Responsibility	
Variable	*F*(1,26)	*Par.* η^2	*F*(1,26)	*Par.* η^2	*F*(1,26)	*Par.* η^2
Role	1.08	.06	2.16	.08	0.42	.02
Alert	65.63****	.72	1.51	.06	52.11****	.67
Order	12.75**	.33	0.31	.01	1.77	.07
Role × Alert	7.77*	.23	0.44	.02	8.52**	.25
Role × Order	4.54*	.15	0.09	.00	1.33	.05
Alert × Order	0.01	.00	1.88	.07	0.26	.01
Role × Alert × Order	5.30*	.17	0.72	.03	0.79	.03

Note. *$p < .05$; **$p < .01$; ***$p < .005$; ****$p < .0001$.

TABLE 5: Actors' and Observers' Mean Subjective Ratings of Q1, Q2, Q3–Q5

	Q1: Subjective Assessment of the Alert Detection Capability		Q2: Subjective Assessment of Human Detection Capability		Q3-Q5: Subjective Assessment of Human Responsibility	
	Less-Accurate system	Accurate system	Less-Accurate system	Accurate system	Less-Accurate system	Accurate system
Actors	3.7[a] (SD = .3)	6.1[b] (SD = .2)	4.3 (SD = .2)	3.9 (SD = .2)	4.5 (SD = .1)	2.5 (SD = .2)
	3.6[b] (SD = .3)	5.3[a] (SD = .3)				
Observers	5.0[a] (SD = .2)	5.9[b] (SD = .3)	4.4 (SD = .2)	4.2 (SD = .2)	4.2 (SD = .2)	3.1 (SD = .2)
	3.6[b] (SD = .2)	5.1[a] (SD = .3)				

Abbreviation: SD = standard deviation.
Notes. [a]When examined first; [b]When examined second; SD = standard deviation.

the less-accurate system, the system's and the human's detection sensitivities were low. We compared participants' answers to questions Q1 and Q2, with the less-accurate system, using a two-way mixed ANOVA, with the question type as a within-subjects variable and the order of observing the less-accurate system as a between-subject variable (mean values are presented in Table 5). The actors perceived their own detection capability as significantly higher than that of the less-accurate system, $F(1,27) = 4.5$, $p < .05$; *Par.* $\eta^2 = .14$.

The observers perceived the actors' and the less-accurate system's capabilities as similar when this system was examined first, $p = .17$. When it was examined second, the observers rated its capabilities as significantly lower than those of the actors, $F(1,13) = 7.5$, $p < .05$; *Par.* $\eta^2 = .37$. Similar comparisons of participants' answers to questions Q1 and Q2 with the accurate system revealed that observers evaluated this system's detection capabilities as significantly higher than those of the actors $F(1,27) = 11.16$, $p < .005$; *Par.* $\eta^2 = .29$, and so did the actors, $F(1,27) = 33.50$, $p < .0001$; *Par.* $\eta^2 = .55$

The analysis of questions Q3–Q5 revealed highly significant effects of the type of alert system and the participant's role. Both types of participants saw the actors' responsibility as significantly lower with the accurate system. However, the observers differentiated less between the actor's levels of responsibility with the two systems.

We analyzed the answers to question F1, in which participants compared their reliance on the two systems, with two-way mixed ANOVA, with participants' role as a within-subjects variable and order as a between-subject variable. The order had no significant main or interaction effect. There was a very large difference between actors and observers, $F(1,26) = 13.95$, $p = .001$; *Par.* $\eta^2 = .35$. The actors stated that they relied much more on the accurate alert system ($M = 6.2$, SD = .2), although the observers stated that actors relied on it only slightly more ($M = 5.0$, SD = .3). Hence, again, the observers differentiated less between the actors' levels of responsibility with the two systems.

DISCUSSION

Actors behaved according to the ResQu model's predictions, assuming significantly more responsibility with the less-accurate system than with the accurate system (see H1). The ResQu model predictions of the actual measured responsibility fit the results for the less-accurate system well. With the accurate system, the actors assumed more than optimal responsibility, due to under-reliance on the system. This is in line with previous results from behavioral research in aided detection tasks, in

which users tended to overestimate their own capabilities and under-trust the system, especially when they performed poorly, compared with the system (Bartlett & McCarley, 2017; Douer & Meyer, 2020b; Meyer et al., 2014).

The experimental points in the current experiment were identical to two experimental points in a previous study that included only actors (Douer & Meyer, 2020b). Actors' behavior was consistent, as measured responsibilities and the cutoff differences were almost identical in the two studies, both for the accurate system, $t(58) = .48$, and the less-accurate system, $t(58) = .80$. Hence, we can conclude that in the current study the observers' presence did not affect the actors' behavior.

H2 was fully supported by the results. Due to the substantial difference between the two systems' detection capabilities, both actors and observers realized that the accurate system had significantly better capabilities than the less-accurate system. The accurate system received significantly higher scores when it was examined second. A possible explanation is that after experiencing the less-accurate system first, the accurate system's performance was perceived as much better.

H3 was also fully supported by the results. Matching the actors' measured responsibility, both actors and observers rated the actors' responsibility as significantly lower with the accurate system than with the less-accurate system.

H4 was supported, with the less-accurate system, and not supported with the accurate system. With the accurate system, the system's superior capabilities and contribution were clear, so both actors and observers rated the system and human capabilities and actors' responsibility similarly, rightly assessing the system's capabilities as higher than those of the actors. However, with the less-accurate system, the humans' and system's capabilities were similarly poor, leading to more adverse outcomes and making it more difficult to ascribe responsibility. In this case, the actors perceived system capabilities as significantly inferior to their own, despite both having similar detection sensitivity. This indicates that, differently from their perception of own errors, actors interpreted the abundance of system errors as representative characteristics of the system, with lesser consideration to exogenous probabilistic factors. Conversely, when the less-accurate system was examined first, the observers perceived its capabilities to be somewhat better than those of the actors. In this case, they attributed the abundance of adverse outcomes mainly to the actors' capabilities. When it was examined second, the observers rated its capabilities to be significantly lower. It seems that in the latter case, the observers generated a reference point for a very good system when observing the accurate system first, which biased down their perception of the less-accurate system. The order of experiencing the systems did not affect the actors, maybe because they tended to attribute much lower capabilities to the less-accurate system, regardless if it was examined first or second. Finally, observers differentiated less between the actors' responsibility and level of reliance across the two systems, both immediately after the trials with each system and at the end of the experiment, after experiencing both systems.

To conclude, actors' and observers' subjective perceptions matched the actual difference in system capabilities and the actors' empirical behavior, even when this behavior was not optimal. They correctly assessed that the better system's capabilities exceeded those of the actors and led to lower human contribution. With the less-accurate system, there were actor–observer differences, and possible attribution errors and other biases arose. Lastly, observers differentiated less between actors' responsibility and reliance on the two systems.

Implications

As predicted by the ResQu model, the result shows that better automation, which greatly exceeds human capabilities, may lower the human comparative responsibility and level of involvement. In line with previous behavioral research results, users may feel (correctly) that they make no significant contribution with such superior systems and may attempt to be more involved by interfering more than necessary. In contrast, they may become complacent and less vigilant to take necessary actions (Hassenzahl & Klapperich, 2014; Moray, 2003; Rangarajan et al., 2005;

Smith et al., 1999). Both responses will probably impair the overall performance. One needs to be prepared to deal with these implications on the overall performance and the humans' attitudes toward advanced intelligent systems and their role in them.

Our results showed actors' systematic tendency to attribute adverse outcomes and errors to system characteristics, in a manner that resembled the fundamental attribution error. Observers differentiated less between the two systems when judging human comparative responsibility, even though they monitored the human–automation interaction closely. Possibly, a more distant observer, such as a manager, might have even larger biases, holding users of intelligent systems responsible for adverse outcomes in situations in which they rightly trusted the system (Douer et al., 2020). This may allow system designers to keep humans in the loop to cope with unexpected events, even when humans may be unable to cope with such events. In this case, humans function as "moral crumple zones," to whom outside observers unjustly assign high moral and legal responsibility when the system fails (Elish, 2019; Elish & Hwang, 2015).

Lastly, our results suggest that periodical presentations of human and system performance measures (e.g., sensitivity and specificity) may aid to calibrate subjective assessments of both actors and observers, reducing human biases and attribution errors when interacting with intelligent systems. The ResQu measured responsibility can aid in this, too, by quantifying the actual marginal level of human contributions to the outcomes.

Limitation and Further Directions

The study was conducted in a controlled lab environment, in which actors performed a simple task and received immediate feedback, and observers attentively examined the interaction and the capabilities of both actors and systems. Further work, which we have started (Douer et al., 2020), should expand the research to real-world settings.

The study was limited to Israeli students, all of which were undergraduate students from the Faculty of Engineering. Future work should examine how cultural and educational differences affect the subjective responsibility attribution in human interaction with intelligent systems.

We made defective parts fairly common to obtain stable Hit and FA probabilities while keeping the experiment short. With rare signals, human operators may become complacent, relying on the alert system to let them know when a problem occurs, assuming that "all is well" otherwise. In such cases, the comparative human contribution to outcomes will decrease. Future work should examine human responsibility with rare events and intelligent classification systems.

Future work should also address temporal effects, such as the time required to make a decision and to act, and its implications on the human's tendency to rely on the automation and the corresponding subjective assessments of human responsibility, made by actors and observers. We plan to address this issue, too.

ACKNOWLEDGMENTS

The research is part of the first author's PhD dissertation at the Department of Industrial Engineering at Tel Aviv University. The research was partly funded by Israel Science Foundation grant 2019/19 to the second author. The experimental platform was developed by Samuel C. Cherna.

KEY POINTS

- Subjective perceptions of responsibility, in human interaction with intelligent systems, are complex and depend on the relative system capabilities and the human's role as an active participant or a passive observer.
- In the interaction with advanced intelligent systems with superior capabilities, users may subjectively feel (correctly) that they do not significantly contribute to the outcomes and may attempt to be more involved, interfering more than necessary and impairing the combined performance.
- Observers may differentiate insufficiently between situations in terms of human comparative responsibility and may fail to consider

the causal contribution of the system to adverse outcomes.

- We demonstrate how the properties of the system and the human's role affect subjective responsibility assessments in an experimental environment.

ORCID iD

Joachim Meyer https://orcid.org/0000-0002-1801-9987

REFERENCES

Abbink, D. A., Carlson, T., Mulder, M., de Winter, J. C. F., Aminravan, F., Gibo, T. L., & Boer, E. R. (2018). A topology of shared control systems-finding common ground in diversity. *IEEE Transactions on Human-Machine Systems*, *48*, 509–525. https://doi.org/10.1109/THMS.2018.2791570

Andrews, P. W. (2001). The psychology of social chess and the evolution of attribution mechanisms: Explaining the fundamental attribution error. *Evolution and Human Behavior*, *22*, 11–29. https://doi.org/10.1016/S1090-5138(00)00059-3

Bartlett, M. L., & McCarley, J. S. (2017). Benchmarking aided decision making in a signal detection task. *Human Factors*, *59*, 881–900. https://doi.org/10.1177/0018720817700258

Bregman, D. (2010). Smart home intelligence–The eHome that learns. *International Journal of Smart Home*, *4*, 35–46.

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature*, *538*, 20–23. https://doi.org/10.1038/538020a

Choi, I., Nisbett, R. E., & Norenzayan, A. (1999). Causal attribution across cultures: Variation and universality. *Psychological Bulletin*, *125*, 47–63. https://doi.org/10.1037/0033-2909.125.1.47

Cicirelli, F., Fortino, G., Giordano, A., Guerrieri, A., Spezzano, G., & Vinci, A. (2016). On the design of smart homes: A framework for activity recognition in home environment. *Journal of Medical Systems*, *40*, 200. https://doi.org/10.1007/s10916-016-0549-7

Coeckelbergh, M. (2012). Moral responsibility, technology, and experiences of the tragic: From Kierkegaard to offshore engineering. *Science and Engineering Ethics*, *18*, 35–48. https://doi.org/10.1007/s11948-010-9233-3

Crootof, R. (2015). The killer robots are here: Legal and policy implications. *Cardozo law review*, *36*, 1837–1915.

Cummings, M. L. (2006). Automation and accountability in decision support system interface design. *The Journal of Technology Studies*, *32*, 23–31. https://doi.org/10.21061/jots.v32i1.a.4

Davison, H. K., & Smothers, J. (2015). How theory X style of management arose from a fundamental attribution error. *Journal of Management History*, *21*, 210–231. https://doi.org/10.1108/JMH-03-2014-0073

Docherty, B., Althaus, R. A., Brinkman, A., Jones, C., & Skipper, R. B. (2012). *Losing humanity: The case against killer robots*. Human Right Watch.

Doi, K. (2007). Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Computerized Medical Imaging and Graphics*, *31*, 198–211. https://doi.org/10.1016/j.compmedimag.2007.02.002

Douer, N., & Meyer, J. (2020a). The responsibility quantification model of human interaction with automation. *IEEE Transactions on Automation Science and Engineering*, *17*, 1044–1060. https://doi.org/10.1109/TASE.2020.2965466

Douer, N., & Meyer, J. (2020b). Theoretical, measured and subjective responsibility in aided decision making. Submitted for publication. *arXiv preprint*. https://arxiv.org/abs/1904.13086

Douer, N., Redlich, M., & Meyer, J. (2020). Operator responsibility for outcomes: A demonstration of the ResQu Model. *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 64). SAGE Publications.

Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human-robot interaction. *Engaging Science, Technology, and Society*, *5*, 40–60.

Elish, M. C., & Hwang, T. (2015). *Praise the machine! Punish the human! The contradictory history of accountability in automated aviation*. Intelligence & Autonomy Working Paper, Data & Society Research Institute. https://doi.org/10.2139/ssrn.2720477

Friedman, B. (1995). *"It's the computer's fault": Reasoning about computers as moral agents* [Conference session]. Paper presented at the Conference Companion on Human Factors in Computing Systems, Denver, CO. 226–227.

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin*, *117*, 21–38. https://doi.org/10.1037/0033-2909.117.1.21

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Wiley.

Hart, H. L. A. (2008). *Punishment and responsibility: Essays in the philosophy of law*. Oxford University Press.

Hart, H. L. A., & Honor, T. (1985). *Causation in the law*. Oxford University Press.

Hassenzahl, M., & Klapperich, H. (2014). *Convenient, clean, and efficient? The experiential costs of everyday automation* [Conference session]. Paper presented at the Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational, Helsinki, Finland. 21–30.

Jalalian, A., Mashohor, S. B. T., Mahmud, H. R., Saripan, M. I. B., Ramli, A. R. B., & Karasfi, B. (2013). Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: A review. *Clinical Imaging*, *37*, 420–426. https://doi.org/10.1016/j.clinimag.2012.09.024

Johnson, M., Bradshaw, J. M., Feltovich, P. J., Jonker, C. M., Van Riemsdijk, M. B., & Sierhuis, M. (2014). Coactive design: Designing support for interdependence in joint activity. *Journal of Human-Robot Interaction*, *3*, 43–69. https://doi.org/10.5898/JHRI.3.1.Johnson

Johnson, D. G., & Powers, T. M. (2005). Computer systems and responsibility: A normative look at technological complexity. *Ethics and Information Technology*, *7*, 99–107. https://doi.org/10.1007/s10676-005-4585-0

Lassiter, G. D., Geers, A. L., Munhall, P. J., Ploutz-Snyder, R. J., & Breitenbecher, D. L. (2002). Illusory causation: Why it occurs. *Psychological Science*, *13*, 299–305. https://doi.org/10.1111/j.0956-7976.2002..x

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: An integrative review. *Theoretical Issues in Ergonomics Science*, *8*, 277–301. https://doi.org/10.1080/14639220500337708

Maltz, M., & Meyer, J. (2001). Use of warnings in an attentionally demanding detection task. *Human Factors*, *43*, 217–226. https://doi.org/10.1518/001872001775900931

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, *98*, 224–253. https://doi.org/10.1037/0033-295X.98.2.224

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, *6*, 175–183. https://doi.org/10.1007/s10676-004-3422-1

Meiring, G. A. M., & Myburgh, H. C. (2015). A review of intelligent driving style analysis systems and related artificial intelligence algorithms. *Sensors*, *15*, 30653–30682. https://doi.org/10.3390/s151229822

Meyer, J. (2001). Effects of warning validity and proximity on responses to warnings. *Human Factors*, *43*, 563–572. https://doi.org/10.1518/001872001775870395

Meyer, J., & Lee, J. D. (2013). Trust, reliance, and compliance. In pp. J. D. Lee & A. Kirlik (Eds.), *The Oxford Handbook of cognitive engineering* (pp. 109–124). Oxford University Press.

Meyer, J., Wiczorek, R., & Günzler, T. (2014). Measures of reliance and compliance in aided visual scanning. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *56*, 840–849. https://doi.org/10.1177/0018720813512865

Moray, N. (2003). Monitoring, complacency, scepticism and eutactic behaviour. *International Journal of Industrial Ergonomics*, *31*, 175–178. https://doi.org/10.1016/S0169-8141(02)00194-4

Morgan, T. (1992). Competence and responsibility in intelligent systems. *Artificial Intelligence Review*, *6*, 217–226. https://doi.org/10.1007/BF00150235

Noorman, M., & Johnson, D. G. (2014). Negotiating autonomy and responsibility in military robots. *Ethics and Information Technology*, *16*, 51–62. https://doi.org/10.1007/s10676-013-9335-0

Rangarajan, D., Jones, E., & Chin, W. (2005). Impact of sales force automation on technology-related stress, effort, and technology usage among salespeople. *Industrial Marketing Management*, *34*, 345–354. https://doi.org/10.1016/j.indmarman.2004.09.015

Robinson, D. E., & Sorkin, R. D. (1985). A contingent criterion model of computer assisted detection. *Trends in Ergonomics/Human Factors*, *2*, 75–82.

Rogoff, E. G., Lee, M.-S., & Suh, D.-C. (2004). "Who Done It?" Attributions by entrepreneurs and experts of the factors that cause and impede small business success. *Journal of Small Business Management*, *42*, 364–376. https://doi.org/10.1111/j.1540-627X.2004.00117.x

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. *Advances in Experimental Social Psychology*, *10*, 173–220.

Ross, L. (2018). From the fundamental attribution error to the truly fundamental attribution error and beyond: My research journey. *Perspectives on Psychological Science*, *13*, 750–769. https://doi.org/10.1177/1745691618769855

Scharre, P. (2016). *Autonomous weapons and operational risk*. Center for a New American Security.

Smith, E. R., & Miller, F. D. (1979). Salience and the cognitive mediation of attribution. *Journal of Personality and Social Psychology*, *37*, 2240–2252. https://doi.org/10.1037/0022-3514.37.12.2240

Smith, M. J., Conway, F. T., & Karsh, B. T. (1999). Occupational stress in human computer interaction. *Industrial Health*, *37*, 157–173. https://doi.org/10.2486/indhealth.37.157

Sorkin, R. D. (1988). Forum: Why are people turning off our alarms? *The Journal of the Acoustical Society of America*, *84*, 1107–1108. https://doi.org/10.1121/1.397232

Sorkin, R. D., & Woods, D. D. (1985). Systems with human monitors: A signal detection analysis. *Human–Computer Interaction*, *1*, 49–75. https://doi.org/10.1207/s15327051hci0101_2

Sparrow, R. (2009). Predators or plowshares? Arms control of robotic weapons. *IEEE Technology and Society Magazine*, *28*, 25–29. https://doi.org/10.1109/MTS.2009.931862

van Dongen, K., & van Maanen, P.-P. (2006). Under-reliance on the decision aid: A difference in calibration and attribution between self and aid. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*, 225–229. https://doi.org/10.1177/154193120605000304

van Dyck, C., Frese, M., Baer, M., & Sonnentag, S. (2005). Organizational error management culture and its impact on performance: A two-study replication. *Journal of Applied Psychology*, *90*, 1228–1240. https://doi.org/10.1037/0021-9010.90.6.1228

Vicente, K. J. (1999). *Cognitive work analysis: Toward safe, productive, and healthy computer-based work*. CRC Press.

Vincent, N. A. (2011). A structured taxonomy of responsibility concepts. In *Moral responsibility* (pp. 15–35). Springer.

Nir Douer is a PhD candidate in the Department of Industrial Engineering at Tel Aviv University, Israel. He has an MSc in Operations Research (1991) and an MA in Economics (2001) from Tel-Aviv University.

Joachim Meyer is a professor in the Department of Industrial Engineering at Tel Aviv University, Israel, was on the faculty of Ben-Gurion University of the Negev, Israel, and held research positions at the Technion, Israel Institute of Technology, and at the MIT AgeLab and the MIT MediaLab. He has an MA in psychology and a PhD in Industrial Engineering (1994) from the Ben-Gurion University of the Negev in Beer Sheva, Israel.

Date received: January 4, 2020
Date accepted: June 5, 2020