Normalcy Bias in Search and Rescue Decisions: Reverse Information Bias in a Sequential 3-way Signal Detection Task

Kevin Kapadia[a], Robin Dillon[b], and Richard S. John[a]
[a] University of Southern California, Los Angeles, U.S.A, kevinkap@usc.edu; richardj@usc.edu
[b] Georgetown University, Washington D.C., U.S.A, rld9@georgetown.edu

Abstract: Effective maritime search and rescue (SAR) requires accurately assessing vessel distress, often based on ambiguous and probabilistic signals. This experiment examined how individuals aggregate diagnostic and nondiagnostic signals when making SAR launch decisions, with a particular focus on normalcy bias, i.e., the tendency to fail to take the appropriate actions because a situation is interpreted as reflecting normal rather than abnormal conditions. Over 800 participants were randomly assigned to conditions that varied in the relative cost of false positives and false negatives and in the base rate of vessel distress. Each participant completed four scenarios, with the number of distress and benign scenarios determined by their assigned base rate condition. Participants could request up to eight signals per scenario before committing to a launch or stand-down decision. Despite the option to request up to eight signals, participants exhibited a reverse information bias, requesting an average of three, with more confident individuals and those who perceived a higher base rate of vessel distress tending to request fewer signals. Participants with higher launch rates were in conditions with more severe false-negative penalties, in a high base-rate condition, judging a distress scenario, had lower confidence in their decisions, stated a higher perceived posterior probability of vessel distress, and described an employment history in a water job. Overall performance was about halfway between random guessing and perfection, but only slightly better than an optimal strategy that uses only the base rate and asymmetric error penalties while ignoring all signal information.

1. INTRODUCTION

Failures in emergency response rarely stem from a lack of information or disengagement. Instead, they often arise from how warning signals are interpreted in the moments before a crisis fully materializes. In high-stress situations, decision makers routinely encounter cues that are weak, ambiguous, or inconsistent with prior expectations. A central mechanism underlying this process is normalcy bias, the systematic tendency to interpret ambiguous or atypical warning signals as consistent with normal, non-threatening conditions, thereby reducing their perceived value and delaying protective action [1]. Normalcy bias does not reflect indifference; rather, it reflects a predictable distortion in how signals are interpreted and aggregated in real time.

Normalcy bias has been documented in a wide range of risk domains. In natural disaster contexts, evacuees and officials delay evacuation or mobilization despite credible warnings, especially when early signals are weak or inconsistent with prior experience [2, 3]. In emergency response settings, failures often involve prolonged information gathering in situations where rapid intervention would have been beneficial [4, 5]. In military and security domains, investigations of operational failures repeatedly identify real-time discounting of anomalous signals that conflict with expectations about adversary intent or environmental conditions [6]. In each area, decision makers are not passive but interpret signals in ways that favor assumptions of normalcy, even when the consequences of delayed action are severe.

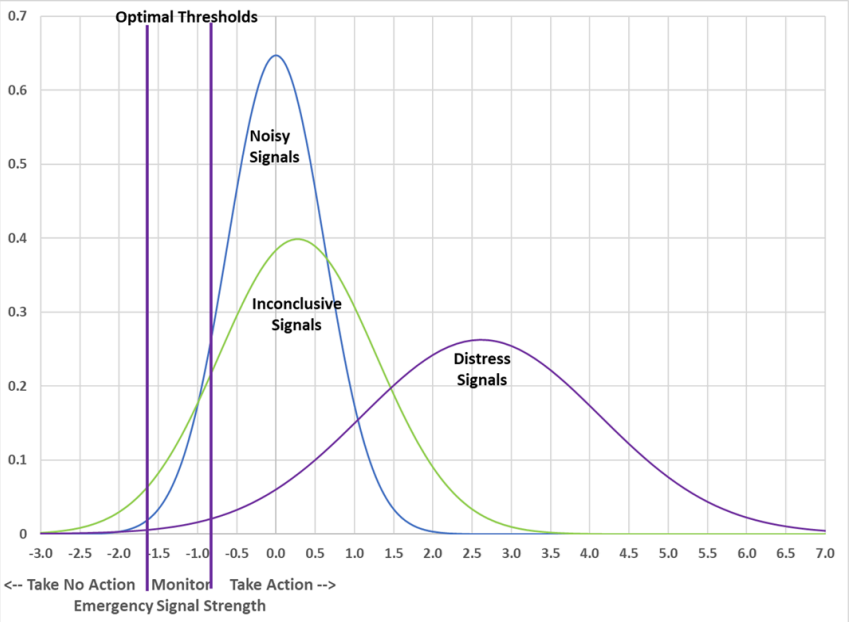
Signal Detection Theory (SDT) provides a normative framework for understanding how normalcy bias can distort decision-making in early risk detection contexts [7]. In the SDT framework, decision-makers must evaluate whether an underlying state of the decision space is hazardous or benign based on noisy, incomplete signals. Since these signals can overlap, any single observation

may plausibly arise from either signal type. The diagnostic value of a signal is captured by its likelihood ratio, the degree to which a signal is more probable under one state than the other. Action is warranted when the cumulative evidence exceeds a decision threshold, which depends on the prior probability of the hazardous state and the relative costs of false positives and false negatives. In this framework, normalcy bias represents a systematic upward shift in decision thresholds, requiring stronger evidence than warranted.

Emergency contexts, such as those faced by USCG search-and-rescue (SAR) decision-makers, are characterized by information arriving in irregular sequences. Typically, the USCG will monitor incoming information related to a possible vessel in distress until enough information is obtained to make one of two possible decisions: 1) Launch a SAR operation, or 2) Conclude there is no vessel in distress and stand down. For this context, an extension of SDT is required that includes an inconclusive category representing the state of monitoring signals. Three-way SDT has been developed and applied in medical decision making [8-11] and forensic science [12]. In the USCG SAR context, information is received, and the situation is monitored until it reaches a threshold that warrants launching a SAR operation or standing down and stopping monitoring of incoming related information.

Figure 1 presents the 3-way SDT framework graphically. USCG decision-makers act to either launch or stand down based on the combined signal strength (T) of signals received so far; thus, the plotted distributions represent those for sequences of multiple signals. In this example, signal sequences received for a distressed vessel follow the purple distress signal curve on the right, while signal sequences when all is well follow the blue noisy signal curve on the left. Signal sequences that are inconclusive follow the green curve in the middle. The purple vertical lines represent thresholds for either standing down (T < left threshold) or launching a SAR mission (T > right threshold). These thresholds are normatively determined from (1) the base-rate (or prior probability) of a distressed vessel, and (2) the relative error penalties for a false positive (launching with no vessel in distress) and a false negative (not launching when there is a distressed vessel).

Figure 1. 3-Way Signal Detection Framework for USCG SAR Decision Making



Normalcy bias can occur from one of three errors in this context: (1) misperception or misaggregation of signal strengths, (2) shift of the SAR launch threshold to the right, requiring more information than needed to launch base-rate misperception, and (3) shift of the stand down threshold to the right, resulting in standing down and ending signal monitoring before enough information to launch a SAR is obtained/misperception of error penalties. Shifts in either threshold can result from neglect or

misapplication of relevant base rates or relative error penalties, such as attenuating the prior probability of a distressed vessel or attenuating the relative cost of a false negative compared to a false positive.

Because of the nature of SAR missions, a delay in deciding to launch can lead to failure. Thus, distress signal sequences that are discounted or diluted by noisy signals will result in a delay in reaching the SAR launch threshold, resulting in normalcy bias. Likewise, neglect of a high base rate of distressed vessels will result in a threshold for launching that is too far to the right, requiring a stronger combination of distress signal sequences to launch. Similarly, ignoring asymmetric error penalties, where false negatives are many times more costly than false positives, will also result in a sub-optimal threshold requiring a stronger sequence of distress signals to launch. Another source of normalcy bias is setting the stand-down threshold too close to the launch threshold, narrowing the window to continue monitoring signals and increasing the risk of standing down prematurely. Standing down is an even more extreme form of normalcy bias, in which a SAR launch is not simply delayed but precluded.

Several cognitive mechanisms can contribute to misperception and misaggregation of signals and base rates. Base rate neglect leads decision-makers to disregard prior probabilities when evaluating new information [13, 14], yielding thresholds that require unusually strong confirmatory evidence before hazardous states are recognized. Conservatism in updating beliefs causes posterior probabilities to revise too slowly as evidence accumulates [15]. Averaging heuristics lead individuals to treat nondiagnostic signals as if they actively counter diagnostic evidence, rather than recognizing them as largely uninformative [16, 17]. Anchoring and order effects distort sequential processing, with early signals anchoring later interpretations and recent information being evaluated disproportionately higher [18, 19]. Information bias refers to the tendency of decision makers to continue collecting more and more data in uncertain contexts, beyond what is needed to take effective action.

Prior work on normalcy bias has largely focused on retrospective case analyses and post hoc examinations of operational failures [1, 20]. While informative, these accounts cannot isolate the specific cognitive mechanisms through which normalcy bias is most likely to occur. The present experiment addresses this gap using a mixed factorial design replicated across two samples. We aim to understand how participant assignment to a base-rate and scoring-rule condition affects evaluations of the number of signals collected, posterior probabilities of a vessel in distress, and the decision to launch a SAR operation or stand down.

2. METHODS

2.1 Overview

The data presented were collected in two samples. Both samples followed the same procedure outlined below. Sample 1 consisted of undergraduate students from Georgetown University's McDonough School of Business. Sample 2 consisted of undergraduate students enrolled in psychology classes at the University of Southern California. Both samples received course credit and an opportunity to win one of three $25 Amazon gift cards for completing the experiment. The experimental protocol was reviewed and deemed exempt by the Institutional Review Board (IRB) at the University of Southern California (ID: UP-25-00111 on 2/18/25).

Participants in both samples followed the same experimental procedure. They were instructed to role-play a Coast Guard officer deciding whether to launch a SAR operation after a recent hurricane in South Florida. They were informed that the Coast Guard's resources were already stretched thin and that committing to a launch would divert assets from other critical tasks and potentially compromise the ability to respond to higher-priority emergencies.

2.2 Warning Signals and Scenario Design

Each participant completed four scenarios, the order of which was randomized. In each scenario, participants received warning signals from junior officers one at a time. The order of signals in each scenario was randomized. Eight warning signals were used across all scenarios, each available in a diagnostic and nondiagnostic version. Diagnostic signals described atypical or concerning conditions that plausibly elevated the posterior probability of vessel distress (e.g., radar logs showing a small vessel in the area the previous night that was no longer appearing on the system). Nondiagnostic signals conveyed a similar context without affirmative evidence of risk (e.g., no record of vessel activity on radar logs in the area during the past 24 hours). Each diagnostic signal was thus paired with a matched nondiagnostic counterpart, holding scenario context constant while varying evidentiary value. Signals were selected from a larger candidate pool based on item response theory (IRT) analyses conducted in a prior calibration study. Table 1 presents the warning signal text for each of the eight signal-label pairs.

Table 1: Warning Signal Text for Signal Labels

Signal Label	Diagnostic	Nondiagnostic
Debris	There is small debris that could be life jackets, cushions, or a cooler visible in the area.	There is no small debris that could be life jackets, cushions, or a cooler visible in the area.
Radio	Recent radio traffic from marinas mentioned unaccounted-for vessels in the region, though no specific details were provided	There has been no recent radio traffic from marinas reporting unaccounted for vessels within the last 6 hours
Flare	There are faint scorch marks on the debris, possibly caused by contact with a handheld distress flare.	There are no visible marine flares indicating distress.
Radar	Radar logs showed a small vessel in the area the previous night, but it is no longer appearing on the system.	There was no record of vessel activity on radar logs in the area during the past 24 hours.
Move	A nearby observer reported that they thought they saw something move near the debris but couldn't confirm exactly what it was.	Observers in the area said they had not seen any movement near the debris.
Growth	The debris appears clean and free of marine growth, suggesting it may have entered the water recently.	The debris shows signs of prolonged exposure, barnacles, algae, or weathering, indicating it may have been drifting for some time.
Call	There was a single call earlier in the day that had referenced a possible person in the water, though it couldn't be verified.	There were no calls recently (in the last 6 hours) that referenced any possible people in the water.
Sheen	A faint sheen on the water was visible near the debris, possibly consistent with oil or fuel.	There is no oil spill or fuel sheen visible near the debris.

Each scenario was either a distress event or a benign event. Signals presented in each scenario were linked probabilistically to the underlying state, and the normative response was to launch in distress scenarios and to stand down in benign scenarios. However, because diagnostic signals could still appear in benign scenarios and nondiagnostic signals in distress scenarios, both types of scenarios remained ambiguous.

2.3 Procedure

Participants were randomly assigned to one of nine conditions in a 3 (Scoring Rule) x 3 (Base Rate) between-subjects design. The scoring rule manipulation determined the cost associated with each type of decision error. In the symmetric condition, false negatives (FN: failing to launch when a vessel was in distress) and false positives (FP: launching when no vessel was in distress) were penalized equally (FN = -100 / FP = -100; Symmetric). In the remaining conditions, the penalties were asymmetric, either emphasizing the cost of missing a distress event (FN = -150 / FP = -50; FN Severe) or the cost of an unnecessary launch (FN = -50 / FP = -150; FP Severe). All participants began with 500 points, with penalties applied for each incorrect decision. The base rate condition specified the prior probability that any given vessel was in distress as 25%, 50%, or 75%. Participants were explicitly informed of their assigned base rate at the beginning of each scenario. The number of distress and benign scenarios was determined by the base-rate condition. Those in the 25% condition completed one distress and three benign scenarios; those in the 50% condition completed two of each; and those in the 75% condition completed three distress and one benign scenario. For analyses involving scoring rule and base rate as predictors, custom polynomial contrasts were applied. The scoring rule was coded to compare the two asymmetric penalty conditions (FN Severe and FP Severe) against the

symmetric condition, and then to compare the two asymmetric conditions against each other. Base rate was coded to compare the two extreme conditions (25% and 75%) against the middle condition (50%), and then to compare the two extreme conditions against each other.

At the start of each scenario, participants received a signal randomly drawn from the eight signal labels, with diagnosticity determined probabilistically by scenario type: 80% diagnostic in distress scenarios and 80% nondiagnostic in benign scenarios. After each signal, participants could choose to launch a SAR, stand down, or continue evaluating and request an additional signal, up to a maximum of eight. Participants who received all eight signals were required to make a final launch-or-stand-down decision at that point. Following their final decision, participants rated their confidence in that decision on a seven-point scale from "Not at all confident" (0) to "Completely confident" (6) and estimated the (posterior) probability that the vessel was in distress on a scale from 0 to 100. These ratings concluded the scenario, after which participants proceeded to the next one.

After completing four scenarios, participants were provided with their final score and feedback indicating which scenarios they had evaluated correctly. They completed the experiment by answering demographic questions and water-related experiences. Table 2 summarizes the demographic information and water experiences for both samples. The median completion time for Sample 1 was 4.48 minutes (Interquartile Range [IQR] = 3.77 - 5.42), and for Sample 2, it was 4.61 minutes (IQR = 3.39 - 6.79).

A post-scenario manipulation check showed that participants were influenced in the expected direction by the base-rate manipulation. In Sample 1, the expected capsize rate was reported as 25.6% in the 25% base-rate condition, 38.9% in the 50% condition, and 58.5% in the 75% condition. In Sample 2, the expected capsize rate was reported as 32.6% in the 25% base-rate condition, 43.8% in the 50% condition, and 54.9% in the 75% condition.

Table 2: Demographic Information for Both Samples

Variable	Level	Sample 1 N (%)	Sample 2 N (%)
Sample Size		476	326
Gender	Female	222 (45.40%)	196 (60.12%)
	Male	266 (54.40%)	129 (39.57%)
	Other	1 (0.20%)	1 (0.31%)
Race/Ethnicity	White Non-Hispanic	299 (61.15%)	76 (26.38%)
	Asian	162 (33.13%)	121 (37.11%)
	Hispanic	48 (9.82%)	54 (16.56%)
	Black	28 (5.73%)	26 (7.98%)
	Other	12 (2.46%)	39 (21.48%)
Water Experience	Kayaking or canoeing	372 (78.15%)	161 (49.39%)
	Recreational boating	317 (66.60%)	126 (38.65%)
	Scuba diving or snorkeling	295 (61.97%)	138 (42.33%)
	Fishing	298 (62.61%)	134 (41.10%)
	Jet skiing	262 (55.04%)	126 (38.65%)
	Waterskiing or wakeboarding	187 (39.29%)	61 (18.71%)
	Sailing	143 (30.04%)	62 (19.02%)
Ever had a water job?	Yes	85 (17.86%)	34 (10.43%)
Age	Mean (SD)	19.52 (1.44)	20.14 (1.76)

2.4 Verification of Randomization for Condition and Signals Displayed

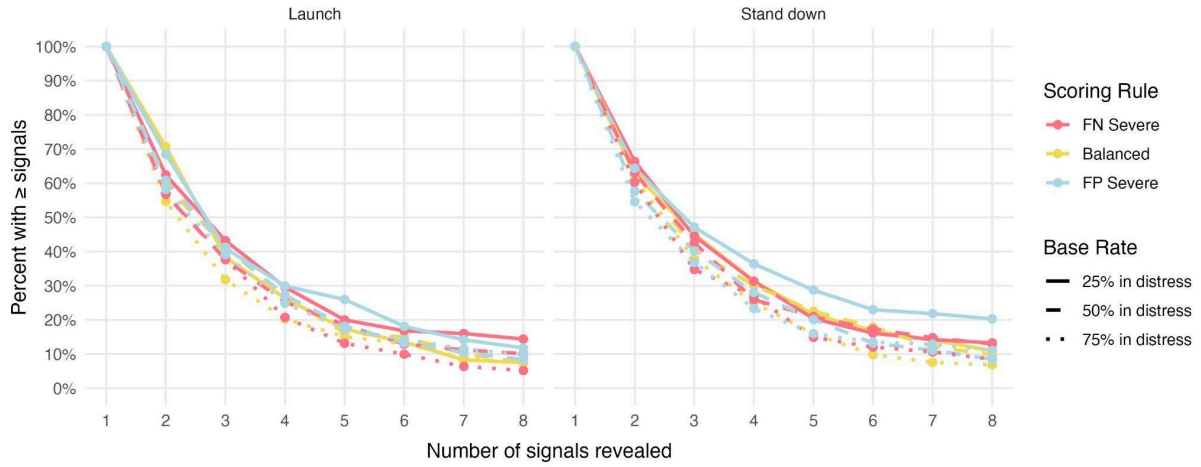
As indicated by Pearson's Chi-square tests of independence, there were no significant differences in the number of participants assigned to each of the nine experimental conditions in both Sample 1 ($X^2(4) = 0.98$, $p = .913$) and Sample 2 ($X^2(4) = 3.18$, $p = .528$). Additionally, Chi-square goodness-of-fit tests confirmed that the eight signal labels were evenly distributed across scenarios in both Sample 1 ($X^2(7) = 3.24$, $p = .862$) and Sample 2 ($X^2(7) = 2.85$, $p = .899$).

3. RESULTS

3.1 Number of Signals Collected

Despite the absence of a penalty for gathering additional information, participants across both samples chose to gather relatively few signals. Because one signal was always presented by default, participants could request up to seven additional signals. In Sample 1, they requested an average of 2.05 additional signals, for a total of 3.05 collected; in Sample 2, they requested an average of 1.57 additional signals, for a total of 2.57 collected. Figure 2 displays the cumulative percentage of participants who collected at least *k* signals by launch decision, scoring rule, and base rate condition. Stopping behavior was heavily front-loaded in both decisions, with the majority of participants making a launch or stand-down decision after requesting only one or two additional signals.

Figure 2. Decumulative Number of Signals Collected by Launch Decision, Scoring Rule, and Base Rate Condition



To examine whether this stopping behavior varied by experimental conditions, confidence in the launch decision, posterior probability of vessel distress, and demographic information, we estimated a Bayesian censored negative binomial mixed-effects model using the *brms* package in R to predict the number of signals gathered. Participants who reached the eight-signal maximum were treated as right-censored using the cens() function. All continuous predictors were standardized, and scoring rule and base rate conditions were contrast-coded as described in the methods. A random intercept for participant was included to account for within-person correlation across scenarios. The model used weakly informative priors: a Normal(1, 0.5) prior on the intercept, reflecting the expectation that participants would collect approximately two to three signals on average ($\log(2.86) \approx 1.05$), Normal(0, 0.3) priors on all fixed effects, encoding the expectation that condition differences would be small on the log scale, an Exponential(2) prior on the random intercept standard deviation, and a Gamma(4, 0.5) prior on the negative Binomial shape parameter. The model was fit with 4 chains of 16,000 iterations each (8,000 warmup), yielding 32,000 post-warmup draws. All Rhat values equaled 1.00, and bulk effective sample sizes exceeded 10,000 across all parameters.

The Bayesian mixed effects model is summarized in Table 3. The model results indicated participants collected fewer signals in Sample 2 (IRR = 0.81, 95% CI = [0.72, 0.90]), fewer signals when more confident (IRR = 0.95, 95% CI = [0.92, 0.98]), and more signals in the 75% base rate condition (IRR = 1.11, 95% CI = [1.04, 1.18]) compared to the 25% condition. The marginal R^2 was .04, the conditional R^2 was .63, and the ICC was .71, indicating that the majority of the variance in the collected signals was attributable to stable individual differences rather than to condition assignment. Overall, these results suggest that although some experimental factors influenced information

gathering, participant-level differences accounted for substantially more variation in the number of signals collected.

Table 3. Bayesian Censored Negative Binomial Mixed Effects Model Predicting Number of Signals

Predictor	Incidence Rate Ratio (IRR)	95% CI
Intercept	**2.80**	**2.51 – 3.11**
Sample [Sample 2]	**0.81**	**0.72 – 0.90**
Scoring Rule Contrast 1 [FN and FP Severe vs Balanced]	1.02	0.95 – 1.10
Scoring Rule Contrast 2 [FP Severe vs FN Severe]	0.98	0.92 – 1.04
Base Rate Condition Contrast 1 [50% vs 25% and 75%]	1.01	0.94 – 1.08
Base Rate Condition Contrast 2 [25% vs 75%]	**1.11**	**1.04 – 1.18**
Scenario Type [Distress]	1.00	0.94 – 1.07
Confidence	**0.95**	**0.92 – 0.98**
Posterior Probability	0.98	0.94 – 1.02
Number of Water Experiences	0.99	0.93 – 1.04
Water Job [Yes]	1.03	0.89 – 1.19
Gender [Male]	1.02	0.93 – 1.13
FinalActionStanddown	1.01	0.94 – 1.09
Scoring Rule Contrast 1 * Base Rate Condition Contrast 1	1.05	0.95 – 1.16
Scoring Rule Contrast 2 * Base Rate Condition Contrast 1	0.96	0.88 – 1.05
Scoring Rule Contrast 1 * Base Rate Condition Contrast 2	1.02	0.93 – 1.11
Scoring Rule Contrast 2 * Base Rate Condition Contrast 2	0.99	0.92 – 1.06

Note: Bolded rows indicate predictors with a significant incident response ratio (IRR).

3.2 Predicting Launch Decision from Experimental Conditions, Confidence, Posterior Probability, and Demographic Information

Participants' decision to launch also varied considerably across conditions. Participants launched in 45.33% of scenarios in Sample 1 and in 50.46% of scenarios in Sample 2. Participants launched in 34.62% of scenarios in the 25% base-rate condition, 47.18% in the 50% condition, and 60.51% in the 75% condition. Participants launched in 51.87% of scenarios in the FN Severe scoring rule condition, 45.74% in the Balanced, and 44.68% in the FP Severe. Finally, participants launched in 24.10% of benign scenarios and in 70.79% of distress scenarios.

To examine whether the experimental conditions, confidence in the launch decision, posterior probability of vessel distress, and demographic information affected participants' decisions to launch a SAR, we estimated a binary logistic mixed-effects regression model with a random intercept for participants. All continuous predictors were standardized, and the scoring rule and base rate were contrast-coded as described in the methods. The binary logistic mixed effects model is summarized in Table 4.

The model found participants were more likely to launch in the FN Severe scoring condition than the FP Severe (OR = 1.35, 95% CI = [1.16, 1.58]), in the 75% base rate condition than 25% (OR = 1.33, 95% CI = [1.13, 1.57]), in distress scenarios (OR = 6.49, 95% CI = [5.10, 8.25]), and when they perceived the posterior probability of a vessel being in distress as higher (OR = 7.29, 95% CI = [6.01, 8.85]). Participants were less likely to launch if they were more confident in their decision (OR = 0.64, 95% CI = [0.56, 0.73]) or had ever had a water job (OR = 0.64, 95% CI = [0.44, 0.93]). The marginal R^2 was .59, the conditional R^2 was .69, and the ICC was .23. Overall, launch decisions were influenced by a range of factors spanning experimental conditions, scenario characteristics, perceived evidence, and participant background.

Table 4. Binary Logistic Mixed-Effects Regression Model Predicting Launch Decision

Predictor	Odds Ratio (OR)	95% CI	p-value
Intercept	**0.28**	**0.22 – 0.37**	**<0.001**

Sample [Sample 2]	1.13	0.86 – 1.49	0.380
Scoring Rule Contrast 1 [FN and FP Severe vs Balanced]	1.15	0.96 – 1.38	0.127
Scoring Rule Contrast 2 [FP Severe vs FN Severe]	**1.35**	**1.16 – 1.58**	**<0.001**
Base Rate Condition Contrast 1 [50% vs 25% and 75%]	1.18	0.98 – 1.41	0.073
Base Rate Condition Contrast 2 [25% vs 75%]	**1.33**	**1.13 – 1.57**	**0.001**
Scenario Type [Distress]	**6.49**	**5.10 – 8.25**	**<0.001**
Confidence	**0.64**	**0.56 – 0.73**	**<0.001**
Posterior Probability	**7.29**	**6.01 – 8.85**	**<0.001**
Number of Water Experiences	1.03	0.90 – 1.18	0.661
Water Job [Yes]	**0.64**	**0.44 – 0.93**	**0.020**
Gender [Male]	1.01	0.78 – 1.31	0.914
Scoring Rule Contrast 1 * Base Rate Condition Contrast 1	0.90	0.70 – 1.16	0.426
Scoring Rule Contrast 2 * Base Rate Condition Contrast 1	0.85	0.68 – 1.05	0.140
Scoring Rule Contrast 1 * Base Rate Condition Contrast 2	1.00	0.80 – 1.25	0.983
Scoring Rule Contrast 2 * Base Rate Condition Contrast 2	1.02	0.85 – 1.23	0.833

3.3 Confidence of Launch Decisions

Participants were generally neutral in their confidence in their launch decisions. On a scale from 0 to 6, participants in Sample 1 had a mean of 2.81 (SD = 1.45), and those in Sample 2 had a mean of 2.79 (SD = 1.36). Participants in the 25% base rate condition had a mean confidence of 2.98 (SD = 1.41), in the 50% base rate condition had a mean of 2.78 (SD = 1.39), and in the 75% base rate condition had a mean of 2.64 (SD = 1.41). Participants in the FN Severe scoring rule condition had a mean confidence of 2.76 (SD = 1.36), in the Balanced condition had a mean of 2.83 (SD = 1.49), and in the FP Severe condition had a mean of 2.81 (SD = 1.39). Finally, participants in benign scenarios had a mean confidence of 2.94 (SD = 1.44), and in distress scenarios, participants had a mean of 2.66 (SD = 1.36).

A linear mixed effects model with the experimental conditions predicting confidence revealed no significant main effect for sample ($F(1, 792.00) = 0.42$, $p = .519$), a significant main effect for base rate ($F(2, 825.33) = 3.12$, $p = .045$), no significant main effect for scoring rule ($F(2, 792.00) = 0.16$, $p = .849$), a significant main effect for scenario type ($F(1, 2405.00) = 27.22$, $p < .001$), and no significant interaction between base rate and scoring rule ($F(4, 792.00) = 0.79$, $p = .530$).

Pairwise comparisons were conducted using estimated marginal means (EMMs) via the *emmeans* package, which computes model-based means for each condition averaged over the other predictors in the model, with Tukey correction applied to adjust for multiple comparisons. For the scenario manipulation, participants reported higher confidence in benign scenarios than in distress scenarios. For the base rate manipulation, pairwise contrasts revealed that participants in the 25% condition reported significantly higher confidence than those in the 75% condition ($t(861) = 2.48$, $p = .036$). The remaining contrasts between the 25% and 50% conditions ($t(808) = 1.49$, $p = .292$) and between the 50% and 75% conditions ($t(808) = 0.96$, $p = .601$) were not significant.

3.4 Participant's Rating for Posterior Probability of Vessel in Distress

Participants' posterior probabilities varied considerably across conditions, ranging from 0 to 100. Participants in Sample 1 had a mean posterior probability rating of 42.41 (SD = 26.24), and participants in Sample 2 had a mean of 46.71 (SD = 26.52). Participants in the 25% base rate condition had a mean posterior probability rating of 34.43 (SD = 25.40), in the 50% condition had a mean of 45.91 (SD = 25.31), and in the 75% condition had a mean of 52.38 (SD = 25.30). Participants in the FN Severe scoring rule condition had a mean posterior probability rating of 44.03 (SD = 25.96), in the Balanced condition had a mean of 44.67 (SD = 27.12), and in the FP Severe condition had a mean of 43.80 (SD = 26.25). Finally, participants in benign scenarios had a mean posterior probability rating of 33.59 (SD = 25.29), and in distress scenarios had a mean of 54.75 (SD = 23.12).

A linear mixed effects model with the experimental conditions predicting posterior probabilities revealed a significant main effect for sample ($F(1, 792.00) = 23.94$, $p < .001$), a significant main effect for base rate ($F(2, 846.22) = 23.38$, $p < .001$), no significant main effect for scoring rule ($F(2, 792.00) = 0.97$, $p = .381$), a significant main effect for scenario type ($F(1, 2405.00) = 564.53$, $p < .001$), and no significant interaction between base rate and scoring rule ($F(4, 792.00) = 0.60$, $p = .662$).

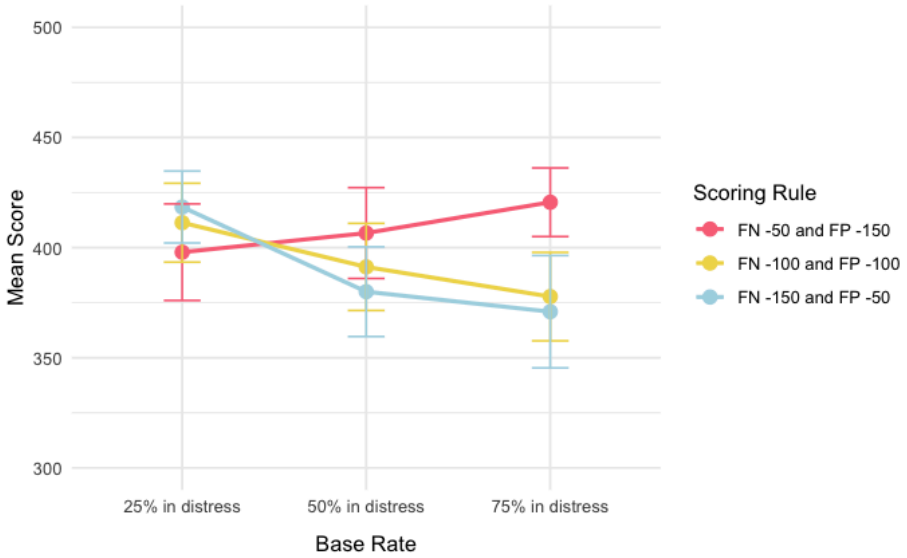
Pairwise comparisons using EMMs found that, for scenario type, participants reported higher posterior probabilities for distress scenarios than for benign scenarios. For example, participants reported higher posterior probabilities in Sample 2. For the base rate, pairwise contrasts revealed that participants in the 25% condition reported significantly lower posterior probabilities than those in the 50% ($t(819) = -5.09$, $p < .001$) and 75% ($t(906) = -6.50$, $p < .001$) conditions. The contrast between the 50% and 75% conditions was not significant ($t(819) = -1.43$, $p = .326$).

3.5 Participant's Accuracy and Score in Scenarios

Overall, participants' scores ranged from the perfectly accurate score (500) to the achievable random-guessing score (300). Out of 500 possible points, the mean score in Sample 1 was 398.84 (SD = 99.84) and in Sample 2 was 395.40 (SD = 95.53). The mean score in the 25% base rate condition was 408.81 (SD = 93.84); in the 50% condition, it was 392.54 (SD = 94.21); and in the 75% condition, it was 390.40 (SD = 104.60). Finally, participants in the FN Severe condition had a mean score of 388.76 (SD = 105.63), in the Balanced condition had a mean score of 394.57 (SD = 91.38), and in the FP Severe condition had a mean score of 408.48 (SD = 95.66).

A 3-way between ANOVA on final score revealed no significant main effect for scoring rule ($F(2, 792) = 2.97$, $p = .052$), no significant main effect for base rate ($F(2, 792) = 2.85$, $p = .059$), no significant main effect for sample ($F(1, 792) = 0.71$, $p = .401$), and a significant interaction between scoring rule and base rate ($F(4, 792) = 3.60$, $p = .006$). The significant interaction is plotted in Figure 3. In the FP Severe condition, participants scored lower than those in other scoring rule conditions when the base rate was low, but outperformed them as the base rate increased.

Figure 3. Significant Interaction Between Base Rate and Scoring Rule Condition Predicting Final Score



Note: Error bars represent 95% confidence intervals.

The accuracy in selecting the correct launch decision was modest. The percentage of scenarios correctly answered in Sample 1 was 73.95%, and in Sample 2 was 72.47%. The percentage of scenarios correctly answered in the 25% base rate condition was 76.71%, in the 50% condition was 71.57%, and in the 75% condition was 71.56%. The percentage of scenarios correctly answered in the

FN Severe scoring rule condition was 71.91%, in the Balanced condition was 73.64%, and in the FP Severe condition was 74.46%. The percentage of scenarios correctly answered was 75.90% in benign scenarios and 70.79% in distress scenarios.

A binary logistic mixed effects model with likelihood ratio tests predicting scenario accuracy revealed no significant effect of scoring rule ($X^2(2) = 1.59$, $p = .451$), no significant effect of base rate ($X^2(2) = 4.71$, $p = .095$), no significant effect of sample ($\chi^2(1) = 1.56$, $p = .211$), and a significant effect of scenario type ($X^2(1) = 5.65$, $p = .017$). The interaction term between scoring rule and base rate was not significant ($X^2(4) = 4.12$, $p = .390$) and was removed from the final model to evaluate the main effects of scoring rule and base rate.

4. DISCUSSION

4.1 Lack of Signals Collected

Notably, participants gathered remarkably few signals before making launch decisions, suggesting a reverse information bias. Although additional signals could be requested without explicit penalty, participants typically stopped after viewing only two or three signals. Within a 3-way SDT framework, this pattern suggests that participants often reached implicit decision thresholds after relatively little information had been accumulated, rather than continuing to sample additional evidence. The negative association between confidence and information gathering is also consistent with this interpretation, as confidence likely reflected the extent to which accumulated evidence had already crossed a launch or stand-down threshold. Importantly, the results suggest that participants may have had relatively low tolerance for prolonged uncertainty or additional information search. From a normalcy-bias perspective, this pattern is consequential because thresholds for escalation may have remained overly conservative, causing ambiguous or weakly diagnostic signals to be interpreted as consistent with routine conditions.

4.2 Factors Influencing Launch Decision

Importantly, participants made launch decisions in an environment characterized by persistent ambiguity rather than certainty. Distress scenarios could still contain nondiagnostic signals, while benign scenarios could still contain diagnostic signals. As a result, no individual signal perfectly revealed the true state of the scenario. Such environments are particularly conducive to normalcy bias because warning signals can be plausibly interpreted as routine anomalies rather than evidence of an unfolding emergency. Participants were therefore required to determine not only how to interpret individual signals, but also how much ambiguous evidence was sufficient to justify escalation.

A combination of the experimental manipulations, scenario evaluations, and stable individual traits influenced launch decisions. As expected, participants were significantly more likely to launch in distress scenarios than in benign ones, suggesting participants were responsive to the provided signals. Launch decisions were also systematically influenced by the manipulated base rates and scoring-rule conditions, consistent with predictions from SDT that decision thresholds should vary with prior probabilities and the relative costs of errors. Participants were more likely to launch when false negatives were more costly and when the prior probability of distress was higher.

At the same time, the results also suggest that participants did not respond to the experimental conditions in a perfectly calibrated manner. Participants still failed to launch in a substantial proportion of distress scenarios and sometimes launched in benign scenarios, indicating that ambiguity in the signal environment produced overlapping interpretations of normal and abnormal conditions. Such overlap is central to normalcy bias because diagnostic signals may be interpreted as routine rather than as evidence requiring escalation.

The relatively modest ICC further suggests that stable individual differences contributed to launch behavior even after accounting for the experimental manipulations and scenario characteristics. This finding is consistent with the idea that individuals may differ systematically in their implicit

thresholds for escalation, tolerance for ambiguity, or interpretation of uncertain signals. Participants with prior water-job experience were also less likely to launch, potentially reflecting greater familiarity with anomalous maritime conditions and a deeper understanding of both the true distress rate and the resources required to launch a SAR.

4.3 Relationship Between Score and Accuracy

Overall, participants' scores reflected moderately accurate performance. Across conditions, mean scores were substantially below the maximum possible score of 500 but above the score expected from random guessing, i.e., 300. The observed mean score of 397 suggests that participants performed better than chance, but only slightly better than a strategy that uses only base rates and error penalties while ignoring all signals, which would yield an expected score of about 389. Variability in scores suggests large individual differences in the ability to accurately interpret and aggregate signals and to apply appropriate decision thresholds using manipulated base rates and error penalties. Although there was a significant interaction between scoring rule and base-rate condition, this interaction primarily reflects the structure of the task environment. In higher base-rate conditions, participants encountered fewer benign scenarios and therefore had fewer opportunities to incur false-positive penalties. Accordingly, participants in the FP Severe condition performed relatively better as the proportion of distress scenarios increased.

4.4 Limitations

Several limitations should be considered when interpreting the results of this experiment. First, both samples were comprised of students primarily earning course credit for their participation. Additionally, these students likely had little training or experience in SAR situations. Second, the scenarios were entirely hypothetical, lacking the context that real-world decision makers would have. While the scoring rules introduced varied penalty costs, real-world scenarios would likely involve multiple competing priorities that decision-makers would have to balance. Third, while the warning signals presented to participants were validated in a previous study, they were not grounded in a specific context and were provided sequentially. Signals in real-world scenarios may be presented in clusters and at uneven time intervals. Real-world SAR decision environments also contain organizational, social, and operational pressures that may further amplify normalcy bias. Decision-makers often operate with concerns about interagency coordination and accountability for unnecessary deployments. Such pressures may strengthen conservative escalation thresholds beyond those observed in the present experimental setting.

4.5 Conclusions

The present experiment examined how participants in two independent samples collected and evaluated diagnostic and nondiagnostic warning signals across four scenarios in which a vessel was potentially in distress. Participants collected an average of three out of eight possible warning signals, with fewer signals collected significantly predicted by overconfidence and higher prior base rates of vessel distress. Higher launch rates were significantly predicted by a greater cost of false negatives and a higher prior probability of distress. More broadly, the findings suggest that normalcy bias in SAR environments may arise not from the complete disregard of warning signals, but from systematic distortions in how ambiguous evidence is gathered, weighted, and judged sufficient for escalation. Participants appeared to rely on thresholds that were too close together for interpreting signals, thereby terminating information searches prematurely and anchoring decisions to prior expectations even when diagnostic signals were available. These dynamics may contribute to delayed escalation in real-world emergency response settings where early signals are often incomplete and ambiguous. Policymakers should be aware of these biases when designing operator training for high-stakes situations involving first responders such as the US Coast Guard.

References

[1] Webster, C. W. R. (2022). Tragic Sinking of Gloucester's Patriot, The. Arcadia Publishing.
[2] Drabek, T. E. (1986). *Hazard perceptions. In [editors] human system responses to disaster: An inventory of sociological findings* (pp. 319-347). New York, NY: Springer New York.
[3] Quarantelli, E. L. (Ed.). (1998). *What is a disaster?: perspectives on the question*. Psychology Press.
[4] Mileti, D. S., & Sorensen, J. H. (1990). Communication of emergency public warnings. *Landslides*, *1*(6), 52-70.
[5] Weick, K. E. (1995). *Sensemaking in organizations*. Thousand Oaks, CA: Sage Publications.
[6] Lopez, C. T. (2024, January 29). *3 U.S. service members killed, others injured in Jordan following drone attack*. DOD News, accessed at: https://www.war.gov/News/News-Stories/Article/article/3659809/3-us-service-members-killed-others-injured-in-jordan-following-drone-attack/
[7] Swets, J. A., & Pickett, R. M. (1987). *Evaluation of diagnostic systems: methods from signal detection theory.* Academic Press.
[8] Edwards, D. C., and Metz, C. E. (2006). Analysis of proposed three-class classification decision rules in terms of the ideal observer decision rule. *Journal of Mathematical Psychology, 50*, 478–487.
[9] Mossman, D. (1999). Three-way ROCs. *Medical Decision Making, 19*, 78–89.
[10] Scurfield, B. K. (1996). Multiple-event forced-choice tasks in the Theory of Signal Detectability. *Journal of Mathematical Psychology, 40*, 253–269.
[11] Scurfield, B.K. (1998). Generalization of the Theory of Signal Detectability to n-event m-dimensional forced-choice tasks. *Journal of Mathematical Psychology, 42*, 5–31.
[12] Scurich, N. & John, R. S. (2023). Three-way ROCs for forensic decision making. *Statistics and Public Policy, 10*(1).
[13] Bar-Hillel, M. (1984). Representativeness and fallacies of probability judgment. *Acta Psychologica*, *55*(2), 91-107.
[14] Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430-454.
[15] Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology*, *72*(3), 346.
[16] Shanteau, J. (1975). Averaging versus multiplying combination rules of inference judgment. *Acta Psychologica*, *39*(1), 83-89.
[17] Troutman, C. M., & Shanteau, J. (1977). Inferences based on nondiagnostic information. *Organizational Behavior and Human Performance*, *19*(1), 43-55.
[18] Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, *185*(4157), 1124-1131.
[19] Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, *24*(1), 1-55.
[20] Weick, K. E. (1995). *Sensemaking in organizations*. Thousand Oaks, CA: Sage Publications.