HUMAN-ROBOT INTERACTION IN SEARCH AND RESCUE: MODELING VICTIM BEHAVIOUR AND TRUST

Author: Maria Grouev   Supervisors: Dr. Anahita Jamshidnejad, Sven Schoonebeek

TU Delft — Delft University of Technology

Background

- Search and rescue is time-critical: victim survival is ~74% within 24h of a disaster, but drops to ~22% after 72h
- Autonomous robots/drones boost SaR speed and efficiency, and keep human responders out of danger
- Yet human–robot interaction (specifically victim trust in rescue robots) is largely unaddressed in SaR literature
- Trust is dynamic and asymmetric: lost quickly, regained slowly; too little trust causes disuse, too much causes misuse — an appropriate "sweet spot" is needed
- Existing SaR simulations (e.g. the GT-BDI model) give victims trust, but robots stay trust-blind, moving in fixed/random patterns, never estimating or responding to trust

Research Question

Main Question:

How can trust between victims and autonomous rescue robots be computationally modelled and incorporated into a search and rescue simulation to improve evacuation outcomes?

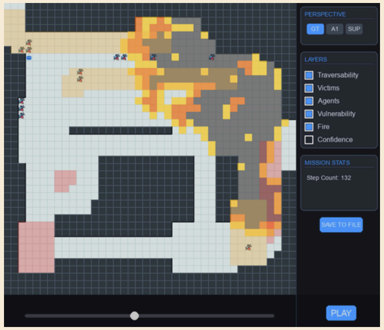
Sub Questions:

1. Modeling: How can victim trust in a rescue robot be modeled as a dynamic variable, and what factors drive its formation, decay, and recovery?
2. Perception: How can a rescue robot estimate a victim's current trust level from observable behavior?
3. Robot strategy: What robot behaviors or strategies most effectively build and sustain victim trust during a rescue interaction?
4. Evaluation: How does the level of victim trust in rescue robots affect evacuation time and survival rate?

Methodology

Testbed

- 2D grid simulation, discrete time steps
- Hazards: fire, vulnerable zones, traversability
- Multiple robots + victims; tracks saved / dead / leftover



Victim Modeling

Trust update: $\tau_i(k+1) = \max\{0, \min\{ \tau_i(k) + \delta_i\, \sigma(k), \tau_i^{max} \}\}$

- $\tau_i(k)$ - victim i's true trust at tick k
- δ_i - personality-based trust sensitivity
- $\sigma(k)$ - input signal = proximity term + danger term (worsening danger penalised 3× harder than improving danger → fast loss / slow gain)
- τ_i^{max} - per-victim trust ceiling

If trust > 0.4 and robot visible → victim defers and follows

Age	Move Prob.	Panicked
Adult	0.900	1.00
Elderly	0.650	0.975
Child	0.525	0.790

Personality	Behavior	Trust Sens. δ_i
Altruist	groups with others, then exits	1.2
Egoist	like altruist; turns selfish after losing a conflict	1.0
Selfish	ignores others, straight to exit	0.6

Agent (robot) modeling:

Trust estimate: $\hat{\tau}(k+1) = \max(0, \min(1, \hat{\tau}(k) + \sigma\_robot \cdot \Delta d(k)))$

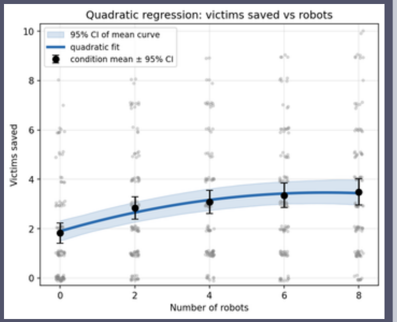
- $\hat{\tau}(k)$ - robot's estimate of victim trust
- $\Delta d(k)$ - change in robot–victim distance (positive = closing in)
- σ_robot - estimate update rate; starts at $\hat{\tau} = 0$ ("assume no trust until shown")
- Estimated from proximity only, the one signal a robot can realistically observe

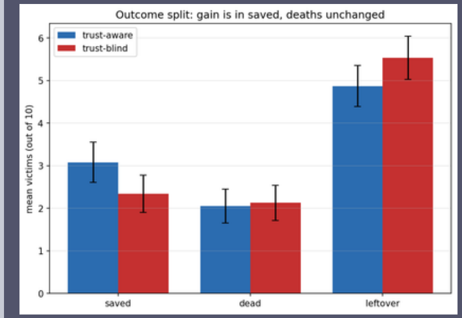
Two modes:

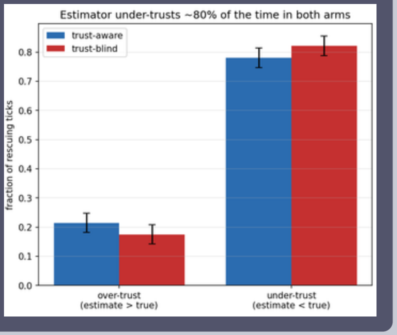
- Approach ($\hat{\tau} < 0.4$): move toward victim to build trust via proximity
- Lead ($\hat{\tau} \geq 0.4$): follow BFS shortest path to exit; waits so it never outpaces the victim

Results

- Trust-aware robot saved significantly more victims: +0.74 on average ($p < 0.001$)
- Gain is in evacuation completion (fewer left stranded), not mortality; deaths unchanged
- Bigger payoff than doubling the fleet: 2 trust-aware robots beat 4 trust-blind robots
- Improvement comes from better positioning/pacing, not higher trust — trust levels were similar across both arms
- Robot consistently under-trusts (~80% of ticks); the safe failure direction: it errs toward staying with the victim rather than abandoning it
- 4 robots chosen for the study: diminishing returns beyond ~4 (quadratic fit, negative term $p < 0.001$)







Conclusion & Future Work

Conclusion:

- Implemented a transparent, bidirectional trust loop: victims hold true trust, robot estimates it and switches modes
- How a robot responds to trust can matter as much as fleet size
- Under-trust bias is the safe failure mode for a safety-critical system

Future work:

- Lower victims' initial trust to study trust building, not just maintenance
- Test varied map sizes/layouts; pre-categorise maps by difficulty to separate trust effects from map noise