SCIENCE TO THE STATION 2026 · KANSAS CITY

Human-AI teaming in fire service operations

Bridging human factors research and frontline fire service decision support

Scott Ramey, MScHQ

PhD Candidate, Systems Design Engineering, University of Waterloo

Founder, The Human Factor · Former Assistant Fire Chief / Deputy Paramedic Chief

Disclosure: Funding received from Defence Research and Development Canada (DRDC), Canadian Safety and Security Program (CSSP)

SCIENCE TO THE STATION · A HEALTH & WELLNESS ALLIANCE

FRED talks

Slides and every source

thehumanfactor.ca/s2s2026-fred

Wednesday, November 4 · 1:05 pm · Imperial Ballroom

MY RESEARCH QUESTION

How do incident commanders calibrate trust in, and appropriately rely on, AI decision support under uncertainty?

Calibrate trust

Match trust to how good the tool actually is.

Rely appropriately

Use it when it is right. Override it when it is wrong.

Under uncertainty

Decisions in seconds, on partial information.

University of Waterloo, Systems Design Engineering · Supervisors Dr. Shi Cao and Dr. Umair Rehman

THE QUESTION

I went looking for the evidence

1,007

records, six databases, 2000 to June 2025

→

32

studies of responders using AI decision support

→

Few

in the fire service. The largest group, 13, came from paramedicine and emergency departments.

Most of what we know about firefighters and AI was learned somewhere else.

Ramey, Masoumi-Ravandi, Rehman and Cao (in preparation). Humans, algorithms, and alarms. University of Waterloo.

SCIENCE TO THE STATION
A HEALTH & WELLNESS ALLIANCE

THE QUESTION

Why the fireground is different

Commanders recognize, then act

Klein’s first study was with fireground commanders. Experts rarely compare options. They recognize the situation and act on the first workable one.

The ironies of automation

The more we automate, the more we depend on the person for the rare, hard case. Endsley argues AI keeps every irony and adds new ones.

Hospital AI is a poor fit

Tools shaped by hospital data and workflows meet decisions made in seconds, on partial information, with gloves on.

AI can strengthen recognition, or interrupt it.

Klein, Calderwood and Clinton-Cirocco 1986 · Bainbridge 1983 · Endsley 2023 · Ramey et al. (in preparation)

FINDING 1 OF 4

Trust is the gatekeeper, and over-trust is the bigger risk

> *“I don’t mind using the system as a tool, but I don’t want to feel like I have to follow its recommendations, especially if something feels off.”*
>
> A paramedic, interview study of 20 prehospital providers

No trust, no use

Trust was the factor most often tied to whether a system got used. Too little and too much both waste it.

Lost faster for machines

People lose confidence in an algorithm faster than in a person after seeing it make the same mistakes.

Experts are not immune

Automation bias hits novices and experts alike, and training alone does not prevent it.

Bai et al. 2025 (participant P20) · Hancock et al. 2011 · Dietvorst et al. 2015 · Parasuraman and Manzey 2010

SCIENCE TO THE STATION
A HEALTH & WELLNESS ALLIANCE

FINDING 2 OF 4

AI helped most when the load was highest

The promise

In a simulated multi-casualty incident, a predictive display improved awareness and decisions, and the benefit was largest under high workload.

The warning, same paper

More information on screen "may easily result in harming SA and performance rather than supporting them."

The value is in the worst ten minutes of your worst call. So is the cost of a cluttered screen.

Parush, Laendler, Springer, Libkind and Jaffee 2020, IEEE CogSIMA · SA = situation awareness

SCIENCE TO THE STATION
A HEALTH & WELLNESS ALLIANCE

FINDING 3 OF 4

Usability decides adoption

Double documentation

“Avoiding double documentation is essential.” Re-entering data kills a tool.

Alarm fatigue

Too many alerts become background noise. Responders asked for fewer, smarter ones.

Hands busy

Paramedics asked for hands-free tools and voice retrieval of protocols.

Experts and novices

Across 12 emergency departments: “excelling with trained users but requiring targeted refinements for novices.”



AI interfaces with PPE, helmets and gloves on: almost no studies.

Nobody has published a test of one in a structure fire.

Bai et al. 2025 · Pinero de Plaza et al. 2025, Int J Med Inform · Ramey et al. (in preparation)

SCIENCE TO THE STATION
A HEALTH & WELLNESS ALLIANCE

FINDING 4 OF 4

Everyone talks about trust. Almost nobody measures it.

8 / 32
named a theory

Few
used NASA-TLX, SAGAT or SUS

0
used a validated trust-in-AI scale

The tools already exist: NASA-TLX · workload | SAGAT · situation awareness | SUS · usability | S-TIAS · trust in AI, 3 items

No controlled studies of responder performance with and without AI. Few field trials. No long-term studies.

Ramey et al. (in preparation) · Hart and Staveland 1988 · Endsley 2021 · Lewis 2018 · McGrath et al. 2025

SCIENCE TO THE STATION
A HEALTH & WELLNESS ALLIANCE

MY RESEARCH

Done, under way, next

How do incident commanders calibrate trust in, and appropriately rely on, AI decision support under uncertainty?

DONE

Systematic review

32 studies. Written.

Scoping review: governing AI

42 sources, none Canadian. Drafted.

UNDER WAY

Study 1: how commanders decide

Hazmat incident commanders, the first 30 minutes, before any AI. Ethics approved. In Recruitment.

NEXT

Study 2: do they catch an AI error?

Simulation, 36 participants. Workload, awareness, trust, overrides. Ethics in review.

Study 3: trust over time

Repeated use over 12 weeks. Ethics in review.

University of Waterloo, Systems Design Engineering · Supervisors Dr. Shi Cao and Dr. Umair Rehman · REB #48583

SCIENCE TO THE STATION
A HEALTH & WELLNESS ALLIANCE

WHERE IT LEADS

Four things fire services can use

1 **Design guidelines**

What AI decision support on the fireground has to do, and must never do.

2 **An evaluation toolkit**

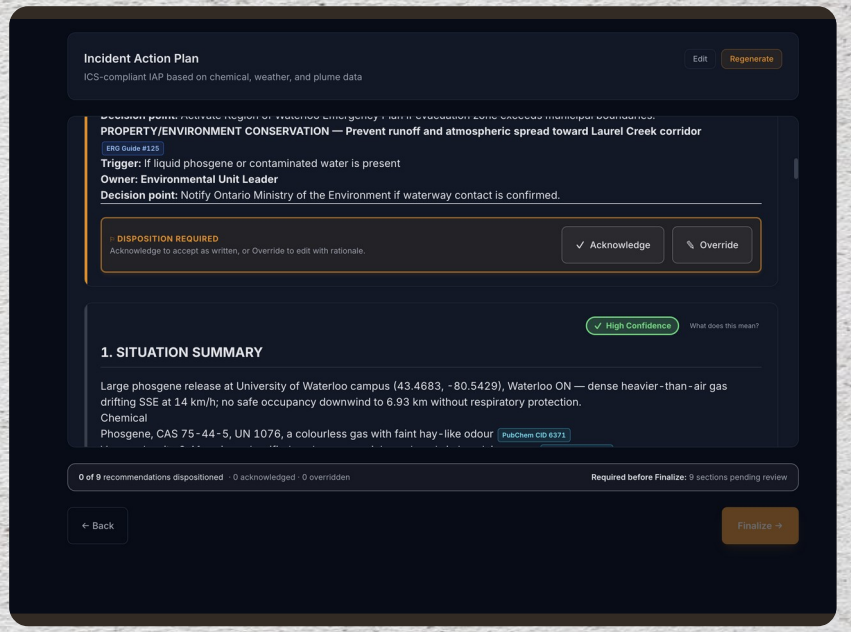
The four measures and a short pilot any department can run before it buys.

3 **A governance template**

What a department's AI policy needs to say, grounded in how services actually govern it.

4 **A decision-support tool built with commanders**

Shaped by Study 1, tested in Studies 2 and 3.



Research prototype, test data. Each recommendation shows its source and how sure it is. Nothing is final until a person acknowledges or overrides it.

Planned outputs of the program · Design informed by Goddard et al. 2012 and Berndt and Herczeg 2019

SCIENCE TO THE STATION
A HEALTH & WELLNESS ALLIANCE

WHAT TO ASK FOR

Until the evidence exists, ask for it

Researchers

Measure with the standard tools, and report four things: who the responders were, how mature the AI was, the outcome measures, and the operational context.

That alone makes the next review a meta-analysis.

Chiefs

Ask any vendor how it was tested, with whom, and what happened to workload, awareness and trust.

If the answer is a demo, you are the experiment.

And for everyone: can your people question, override and correct it, without fear?

Research priorities and reporting items from Ramey et al., systematic review (in preparation)

SCIENCE TO THE STATION
A HEALTH & WELLNESS ALLIANCE

CLOSE

Be the field site, not the afterthought

Thirty-two studies, and almost none about us. That isn't a reason to wait. It's a reason to lead.

Partner with a researcher

Pilot with measurement

Share what you find

Putting Science into Practice · today, 2:45 pm · Trianon E

How to evaluate an AI tool in your own department, with these measures.

Slides and sources

thehumanfactor.ca/s2s2026-fred

Scott Ramey · University of Waterloo, Systems Design Engineering

Sources

Bai et al. 2025. Enhancing prehospital decision-making. BMC Med Inform Decis Mak 25:31. 10.1186/s12911-024-02844-1

Bainbridge 1983. Ironies of automation. Automatica 19(6). 10.1016/0005-1098(83)90046-8

Berndt and Herczeg 2019. Decision support for mass casualty incidents. IFAC-PapersOnLine 52(19). 10.1016/j.ifacol.2019.12.096

Dietvorst et al. 2015. Algorithm aversion. J Exp Psychol Gen 144(1). 10.1037/xge0000033

Endsley 2021. SAGAT and SPAM: a meta-analysis. Human Factors 63(1). 10.1177/0018720819875376

Endsley 2023. Ironies of artificial intelligence. Ergonomics 66(11). 10.1080/00140139.2023.2243404

Goddard et al. 2012. Automation bias: a systematic review. JAMIA 19(1). 10.1136/amiajnl-2011-000089

Hancock et al. 2011. Trust in human-robot interaction. Human Factors 53(5). 10.1177/0018720811417254

Hart and Staveland 1988. Development of NASA-TLX. Advances in Psychology 52. 10.1016/S0166-4115(08)62386-9

Klein et al. 1986. Rapid decision making on the fire ground. Proc HFS 30(6). 10.1177/154193128603000616

Lewis 2018. The System Usability Scale. Int J Hum-Comput Interact 34(7). 10.1080/10447318.2018.1455307

McGrath et al. 2025. Measuring trust in AI: the S-TIAS. Front Artif Intell 8. 10.3389/frai.2025.1582880

Parasuraman and Manzey 2010. Complacency and bias. Human Factors 52(3). 10.1177/0018720810376055

Parush et al. 2020. Predictive situation display for EMS. IEEE CogSIMA. 10.1109/CogSIMA49017.2020.9216126

Perry et al. 2012. Reduced processing decision support interfaces. Decis Support Syst 52(2). 10.1016/j.dss.2011.10.010

Pinero de Plaza et al. 2025. Evaluating RAPIDx AI. Int J Med Inform 196. 10.1016/j.ijmedinf.2025.105810

Ramey and Rehman. Governing AI use in emergency services: a rapid scoping review. Draft.

Ramey, Masoumi-Ravandi, Rehman and Cao. Humans, algorithms, and alarms: a systematic review. In preparation.

Full list with links: thehumanfactor.ca/s2s2026-fred

SCIENCE TO THE STATION
A HEALTH & WELLNESS ALLIANCE