AI Makes The Call. Who Answers For It?

Accountability, Authority, And The Data Infrastructure That Makes Ethical AI Possible In Wildland Firefighting.

Most people working in fire operations have formed a view on AI. Whether that view is broadly positive, cautious, or somewhere in between, it was almost certainly shaped by experiences that have little to do with aerial firefighting: a consumer tool that impressed, a headline that alarmed, a demonstration that felt either compelling or unconvincing. That is a reasonable way to form an initial impression of a new technology. It is a less reasonable basis for procurement decisions, governance frameworks, and accountability structures in a life-safety environment.

This paper is about the gap between those two things. What the science tells us about how humans actually perform alongside automated systems. What operational data from real fire seasons reveals about the reliability of the inputs feeding those systems. And what the industries that have already navigated this challenge learned, mostly at cost, about what accountable AI deployment actually requires.

Contents

The line has already moved.

The wildland firefighting industry did not sit down and decide to adopt AI. It arrived incrementally, through tools that got a little smarter each season. Dispatch systems that started surfacing patterns. Spread models that began updating from satellite feeds rather than manually entered weather data. Alert platforms that moved from notifying after a human decision to notifying in place of one.

At each step the change felt like a natural improvement. Cumulatively, many agencies are now relying on algorithmic recommendations for decisions that were once the exclusive domain of experienced human judgment, without having formally decided that was acceptable, and without a clear framework for who is responsible when something goes wrong.

This is not a criticism of those agencies. It is the same pattern every industry has followed when capable automation arrived faster than the governance frameworks needed to manage it. Aviation went through it. Medicine went through it. Nuclear power went through it. In each case the technology arrived before the accountability structures, and the cost of that sequencing was paid in incidents that should not have happened.

The question for wildland firefighting is not whether AI belongs in the operational environment. It clearly has a role, and that role will grow. The question is whether the industry builds the governance layer before an incident forces it, or after.

> "Many agencies are now relying on algorithmic recommendations for decisions that were once the exclusive domain of experienced human judgment."

TracPlus has provided operational data infrastructure for aerial firefighting at national scale for fifteen years. CAL FIRE, NSW Rural Fire Service, and Australia's national aerial firefighting program are among our closest partners. We record every flight, every drop, every position across more than 638 customers in 44 countries. We are not an AI company. We are a data company, and we have a direct view of how these tools actually perform when aircraft are in the air and decisions are being made under pressure. That perspective is what this paper draws on.

What the science tells us.

The research that matters most here did not come from wildland firefighting. It came from aviation, medicine, and industrial process control, industries that encountered capable automation earlier, studied what it did to human performance, and produced findings that transfer directly to fire operations.

"When humans work alongside automated systems that are reliable most of the time, their ability to critically evaluate those systems degrades."

The central finding is straightforward. When humans work alongside automated systems that are reliable most of the time, their ability to critically evaluate those systems degrades. Not because they stop caring, but because sustained skepticism toward something that is consistently right is cognitively expensive and the brain stops paying that cost. Parasuraman and Manzey's 2010 review of this phenomenon, drawing on decades of research across high-stakes operational environments, found that this over-reliance intensifies under high cognitive load, time pressure, and fatigue. Which is to say it intensifies under exactly the conditions that fire operations produce.

Lisanne Bainbridge named what she called the ironies of automation in 1983, writing about industrial process control but describing something that applies anywhere automation has become capable. The better a system performs, the more the human is reduced to a monitor. The more the human monitors rather than actively operates, the less capable they become of intervening effectively when the system fails. The skill degrades at roughly the rate the system's reliability increases. By the time the system encounters conditions it cannot handle, the human who should be able to take back control may no longer be well-equipped to do so.

Mica Endsley's situational awareness research adds the specific mechanism. She describes three levels of awareness: perceiving what is in the environment, understanding what it means, and projecting what will happen next. AI systems in operational firefighting typically absorb that third level. The spread model generates the forward picture. When humans stop constructing that projection themselves the capacity does not disappear immediately. It fades, gradually and without the person noticing, until the moment it is needed and is no longer reliably there.

Air France 447 is the most documented consequence of this sequence. When the autopilot disengaged over the Atlantic in 2009 the crew were highly trained and experienced. What they could not do was reconstruct the situational picture they needed under the conditions that actually existed: sudden loss of automation, degraded instruments, the middle of the night. Despite 75 stall warnings over four minutes they made inputs that deepened rather than corrected the stall. All 228 people on board died. The investigation found not a failure of individual competence but the predictable outcome of sustained reliance on a system that had been performing the cognitive work the situation now demanded of them.

Aviation did not conclude from this that automation was dangerous. It concluded that automation without appropriate governance was dangerous, and built the training standards, interface requirements, and oversight protocols that followed from that conclusion. That is the model.

One additional dimension worth naming. Trust in AI tends to be bundled rather than specific. Most people's general disposition toward AI, confident or skeptical, was formed through experiences with consumer tools like ChatGPT that share almost nothing with operational fire management systems beyond the category name. A positive experience with a navigation app or a language model does not transfer cleanly to evaluating a fire spread model operating on sensor data with life-safety implications. Neither does a negative one. The evaluation framework most people carry into AI procurement decisions was built from the wrong evidence base, and the industry needs better tools for assessing specific operational systems on their own terms.

The autonomy spectrum and where agencies sit.

Automation researchers have long described a spectrum running from fully manual operation at one end to fully autonomous action at the other. In practical terms, for fire operations, it has four meaningful positions.

At the first, data and tools present information and humans make every decision independently. The system shows, the human decides. Judgment is fully intact and fully owned.

At the second, the system surfaces recommendations. A spread model suggests a resource position. A dispatch tool ranks response options. The human evaluates the recommendation and decides whether to follow it. Decision authority remains with the human but the algorithm is now shaping what options they see, in what order, and with what apparent confidence.

At the third, the system acts unless a human intervenes to stop it. Notifications send, resources are alerted, responses are triggered by default. The human's role has shifted from deciding to monitoring and overriding. The threshold for action is no longer a human choice but a human veto.

At the fourth, the system acts within an operational window without the practical possibility of human intervention. This is rare in wildland firefighting today. It is less rare than it was two years ago.

Most agencies, if asked, would place themselves at the second level. The more useful question is where their tools actually sit based on how they behave, not how they are described. A few questions worth asking honestly.

When your system produces a recommendation, what percentage of the time does someone actively evaluate it against independent judgment before acting on it? Not in principle. In practice, under operational pressure, late in a long shift.

If your alerting or dispatch system encountered a sensor failure or connectivity drop right now, would your operators know immediately? Would they know what the system's last output was based on, and whether that basis was still valid?

If you ended a shift today and someone asked what your AI recommended at a specific moment and why, could you reconstruct that from your operational record?

If the honest answer to any of those questions is uncertain, the agency is likely operating at a higher level of automation than it has formally decided to, without the governance structures that decision would require. That is not a criticism. It is the starting point for building them.

What we have seen.

The science describes a predictable pattern. The operational reality of wildland firefighting data provides the conditions for it.

AI systems are only as reliable as the inputs feeding them. That observation is uncontroversial in principle. In practice it is harder to act on than it sounds, because the gap between what an interface projects and what the underlying data actually shows is invisible by design. A system consuming degraded inputs does not display a warning. It produces outputs. Those outputs look the same whether the data behind them is sound or not. The confidence the interface projects is a function of the system's design, not the quality of its inputs.

"AI systems are only as reliable as the inputs feeding them."

In aerial firefighting, the data feeding operational systems comes from sensors on aircraft operating in demanding physical environments: variable temperatures, vibration, inconsistent connectivity, equipment

that in many cases has not been validated since installation. GPS quality varies with conditions and degrades over the course of a season. None of this is visible to the operator working from the interface. None of it is visible to an AI system consuming the feed.

The question is not whether data quality varies in these conditions. It does, and the physical environment guarantees it will. The question is whether the governance structures around AI deployment are designed to catch that variation before it reaches recommendations that operators trust and act on.

Most current deployments are not designed that way. Validation tends to be retrospective rather than continuous, which means problems surface in post-season analysis rather than mid-season correction. An AI system operating across a full fire season on inputs that were never properly validated has been producing confident-looking outputs on an uncertain foundation from the start.

The infrastructure is capable of producing genuinely useful intelligence. The data flowing through it needs to be actively managed rather than assumed to be sound.

Who is responsible when it goes wrong?

In March 2019, five months after Lion Air Flight 610 went down in the Java Sea killing all 189 people on board, Ethiopian Airlines Flight 302 crashed under nearly identical circumstances six minutes after takeoff. 157 people died. In both cases, the Boeing 737 MAX's Maneuvering Characteristics Augmentation System, known as MCAS, had activated based on a faulty sensor reading and pushed the aircraft's nose down repeatedly. The crews could not overcome it.

The subsequent investigation by the US House Transportation and Infrastructure Committee reviewed 600,000 pages of documents and heard from Boeing and FAA officials over five formal hearings. Its 245-page report found that the crashes were not the result of a singular failure or technical mistake. They were the culmination of faulty technical assumptions, a lack of transparency from Boeing's management, and insufficient FAA oversight. Boeing had not included MCAS in the pilot flight manual. Crews encountered a system actively fighting their inputs without knowing it existed. Boeing had admitted in its own internal documents that emphasising MCAS as a new function would increase certification and training requirements, and proceeded on the basis that this was a cost worth avoiding.

Boeing subsequently settled a criminal charge of conspiracy to defraud the United States for $2.5 billion, admitting to misleading statements and omissions about MCAS.

The accountability question the congressional investigation asked is the one this paper is asking. When an automated system makes decisions that contribute to loss of life, and the chain of responsibility runs from software

design through regulatory approval through operator training to the cockpit, where does accountability sit? The investigation's answer was that it sits across that entire chain, and that the absence of clear governance structures had made accountability invisible until the crashes made it unavoidable.

The lesson is not specific to aviation. It applies wherever capable automation is deployed in a life-safety context without a clear framework for who is responsible for what. The Boeing case illustrates what happens at the extreme end of that failure. The more common version, in firefighting as in other industries, is less dramatic and less visible: decisions made in good faith on the basis of algorithmic recommendations that turned out to be less reliable than the interface suggested, with no clear record of what was recommended, on what data, and what the human decided to do with it.

At the moment an AI-assisted decision is made, is the human genuinely deciding or ratifying a recommendation? The science described in section one tells us those two things increasingly blur as operator familiarity with a system grows. Most agencies have no framework for distinguishing them in the moment and no record that would allow the distinction to be reconstructed afterwards.

The regulatory environment in the United States is not currently helping. The EU AI Act classifies AI used in emergency dispatch and first response as high-risk, requiring pre-deployment conformity assessments, mandatory human oversight mechanisms, and detailed operational logs. The US has moved in the opposite direction. The January 2025 executive order framing AI regulation as a barrier to American competitiveness shifted the burden of verification from regulators back to agencies. This is a reality that will become visible the first time a post-incident investigation seriously examines the role of an algorithmic recommendation in a fatality.

Introducing AI responsibly: the parallel validation principle.

AI adoption in firefighting does not have to be all or nothing, and framing it that way misses the more useful question. It is a continuum of involvement, and the critical question at each point on that continuum is whether increased AI involvement is actually producing better outcomes than the human judgment it is supplementing. That question sounds straightforward. Answering it honestly is harder than most agencies currently have the infrastructure to do.

Aviation did not introduce autopilot by deploying it fully and trusting that it would work. New systems were validated against existing performance baselines before they were given operational authority. The question was always: is this producing better outcomes than what it is replacing, under the conditions we actually operate in, not just the conditions the vendor tested it in. That validation was not a one-time gate. It was continuous, because systems that perform well in one season's conditions may not perform the same way in the next.

The equivalent process for fire agencies is to introduce AI decision support in parallel with real operational scenarios before relying on it. Run the AI recommendation alongside the human decision. Record both. Compare them against outcomes over time. That comparison is how you know whether the tool is actually helping, whether it performs consistently across your specific geography and fuel conditions, and whether there are categories of scenario where it should not be trusted. It is also how you know when a system that was performing well has stopped doing so, which the GPS degradation data in this paper suggests can happen across an entire season without anyone noticing.

That process has a single non-negotiable dependency: a complete, agency-owned operational baseline to measure against. Not data held in a vendor's system, accessible at the vendor's discretion, in formats the vendor controls. The agency's own record of what happened, when,

on what aircraft, under what conditions, with what outcomes. Without that baseline the parallel validation process has nothing to run against. Adoption decisions rest on vendor claims and general impressions rather than evidence from the agency's own operations.

"Operational data infrastructure is not a supporting capability that follows from AI adoption. It is the precondition for AI adoption that can be evaluated, validated, and defended."

This reframes the data question in a way that matters for procurement. Operational data infrastructure is not a supporting capability that follows from AI adoption. It is the precondition for AI adoption that can be evaluated, validated, and defended. An agency that builds that foundation first is in a fundamentally different position when an AI vendor comes to the table: it can assess what is being offered against its own evidence rather than against the vendor's. That asymmetry, in the agency's favor, is what accountable AI deployment actually looks like in practice.

What good looks like.

The industries that have navigated this well did not do so by slowing down automation. They did it by asking harder questions of the systems they were deploying, earlier in the process than felt necessary at the time.

The first question is about explainability, and it is more specific than it sounds. Not whether a system can produce a technical explanation of its outputs, but whether an operator under pressure can understand the reasoning behind a recommendation well enough to evaluate it independently. If the basis for a recommendation cannot be communicated in the time available to act on it, the operator is not deciding. They are ratifying. That question has a practical procurement form: ask any vendor to demonstrate what their system shows an operator when it is uncertain, and what it shows them when the data it is working from is degraded. The answer will tell you more than any accuracy benchmark.

"The governance structures that would make AI deployment in wildland firefighting genuinely accountable are still being built."

Which raises the prior question, because explainability is only meaningful if the underlying data is trustworthy. The findings in this paper describe what happens when it is not: confident-looking outputs produced on inputs that are significantly wrong, with no visible indication at the interface that anything has changed. A system's behavior at the edges of its competence is the relevant test, not its performance under normal conditions. Physics-based models fail in diagnosable ways. Machine

learning systems can produce confident outputs while operating well outside the conditions they were trained on. The question worth asking is not how accurate is your system but what does it show the operator when normal conditions no longer apply.

Both of those questions assume something that cannot be taken for granted: an operational record sufficient to reconstruct what actually happened. After any significant incident the questions that matter are reconstruction questions. What did the system recommend, at what time, on the basis of what inputs, with what confidence, and what did the operator decide. Whether that reconstruction is possible depends entirely on decisions made at procurement. An agency that cannot answer those questions from its own records is carrying an accountability exposure it may not have consciously accepted. The same record that answers post-incident questions is also the baseline that makes ongoing validation possible, which means the governance requirement and the operational improvement requirement are the same infrastructure.

The last question follows directly from Bainbridge. Using a system effectively and maintaining the judgment to override it are different skills, and the second degrades as the first improves. Aviation's response was recurrent manual training as a maintained and assessed competency, not a theoretical fallback. The question for fire agencies is whether their training programs address those two skills separately, and whether override capability is tested under conditions that resemble the ones in which it would actually be needed.

None of these questions have comfortable answers in the current environment. That is the point. The governance structures that would make AI deployment in wildland firefighting genuinely accountable are still being built. Asking the questions now, before an incident forces them, is how that building happens.

Create the foundation before you build on it.

AI will continue to enter wildland firefighting. The market pressure, the genuine utility of the tools in the right conditions, and the pace of development all point in one direction. The question available to agency leadership is not whether but on what terms.

The science reviewed in this paper tells us what happens when capable automation meets under-prepared governance. The pattern is consistent across decades and across industries. It does not require bad intentions or incompetent operators. It requires only that the technology arrive faster than the structures needed to use it accountably. That is the current situation in wildland firefighting.

The operational data tells us something the science cannot: the specific vulnerability in this industry. Inputs that are less reliable than the interface suggests. Degradation that builds across seasons without detection. A gap between what the data shows and what decision makers believe it shows. That is not a theoretical risk. It is a documented finding from two years of systematic analysis across half a million operational records.

"Agencies need to own their operational data in a form that supports reconstruction and parallel validation."

Three things follow from this.

Agencies need to locate themselves honestly on the automation spectrum, not where vendor materials describe their tools, but where those tools actually sit based on how they behave when conditions degrade, connectivity drops, or a sensor fails.

Agencies need to own their operational data in a form that supports reconstruction and parallel validation. Not because something has gone wrong, but because the ability to know whether AI is actually improving outcomes requires it. That data belongs to the agency. It should be held in formats the agency controls, and the contract should say so.

The accountability question needs to be answered before an incident forces the answer. When an AI-assisted decision contributes to a bad outcome, the chain of responsibility running from the developer through the agency to the commander in the field does not have a clear governance framework in the current US environment. Building that framework is the work that needs to happen now, across procurement decisions, training design, and operational protocol.

An AI model is a temporary tool. It will be updated, replaced, or retired. The operational data record is the permanent record of an agency's professional judgment. It is the evidence base for post-incident review, budget justification, and the ongoing assessment of whether the tools the agency relies on are earning the trust placed in them.

Build that foundation first. Everything else follows.

About TracPlus.

TracPlus is the leading provider of operational intelligence platforms for mission-critical aviation operations. We provide the operational data infrastructure for aerial firefighting at national scale, serving 638 customers across 44 countries including CAL FIRE, NSW Rural Fire Service, and Australia's national aerial firefighting program. We have recorded over 6.5 million flight hours and have been the only commercial provider of all-of-country aerial firefighting solutions in Australia and New Zealand since 2010. We believe the integrity of operational data is the precondition for any accountability framework that holds when it matters most.

About the author.

Todd O'Hara is CEO of TracPlus, leading the team that helps government agencies and enterprises turn complex aerial firefighting operations into reliable, defensible data. Todd's career sits at the intersection of aviation, safety, and mission-critical technology. He has held leadership roles in both tech and aerial operations and brings hands-on experience as a pilot and flight instructor across helicopters, fixed-wing aircraft, and gliders.