

From AI features to AI foundations: The 2026 imperative

Why 2026 is the year AI stops being a feature and starts being the infrastructure. By Fernando Amendola and Thierry Fautier



Three years of AI in Media & Entertainment have produced a clear pattern: 2024 was the year of discovery, 2025 the year of experimentation, and 2026 must be the year of courageous implementation. What separates the winners from the rest now is not access to better technology, it is the organisational will to rebuild around it.

The era of bolt-on AI features, siloed pilots, and shiny demos is over. What is emerging in its place is something far more consequential: AI as invisible,

essential infrastructure. Not a product, not a department, not an offering, but the actual driver of future business.

This is not a forecast for the timid. The industry is at an inflection point where the decisions made in the next twelve months will determine who controls the viewer relationship for the rest of the decade. We have identified five fronts where that battle will be won or lost.

1. Bespoke TV: The End of the Monolithic Audience

The single most disruptive idea in media

right now is also the most obvious: audiences are not monoliths, and the industry has been treating them like one for decades. In 2026, that comfortable fiction will start to collapse rapidly. We are entering the era of “Bespoke TV”, where content is hyper-tailored by language, location, cultural register, and advertising context, simultaneously, and at scale. AI-driven localization is the fastest-growing operational segment in the industry, moving decisively beyond subtitle tracks into deep phoneme-based voice synthesis and cultural adaptation that survives high-compression

“In 2026, the industry will stop treating audiences as monoliths. We are entering the era of Bespoke TV: my language, my location, my ads.”

distribution cycles. The long tail of global content, markets where traditional dubbing economics have always failed, is suddenly addressable.

But radical personalization will meet its first real structural wall: what we call the Privacy-Personalization Paradox. GDPR in Europe and rapidly evolving US privacy legislation are not going away. The operators who win in this environment will not be those who collect the most data – they will be those who can deliver a genuinely localized, genuinely personal experience without overstepping the boundaries that regulators and increasingly privacy-literate audiences will enforce. This is a design problem as much as a technical one, and most of the industry is not ready for it.

“In 2026, the industry will stop treating audiences as monoliths. We are entering the era of Bespoke TV: my language, my location, my ads.”

2. The Button Disappears: Autonomous Agents Replace the Copilot

In 2025, the industry fell in love with Copilots, AI tools that waited patiently for a human to click a button, approve a decision, or prompt the next action. They were impressive. They were also fundamentally limited by the assumption that a human needed to be in the loop for every step.

In 2026, the button begins to

disappear. The shift toward Autonomous Supply Chains is not a marginal upgrade, it is a structural redesign of how media operations work. These agents do not wait for commands. They detect events – a rights window opening, a competitor’s content spike, a live stream anomaly – and trigger cascades of action: automated rights verification, intelligent clipping, multi-platform distribution, all governed by predefined business rules and high-level human intent. Two emerging standards are making this interoperable: the Model Context Protocol (MCP), which gives agents a universal grammar to communicate across siloed operational systems, and the Ad Context Protocol (AdCP), its commercial-layer counterpart, enabling advertiser and operator agents to negotiate and transact directly without human intermediaries.

The human role does not disappear; it transforms. The broadcast engineer or media operations professional of 2026 is not a button-clicker. They are an Editor-in-Chief of an autonomous supply chain, defining the rules of engagement, reviewing exceptions, and auditing outcomes. This requires a fundamentally different skill set and a fundamentally different relationship with technology. Organizations that have not begun this transition are already behind.

“The lesson of the last two years is clear: AI features don’t scale. AI workflows do.”

3. Infrastructure Sovereignty: Reclaiming the Compute Tax

There is a number that does not appear in most AI transformation case studies, and that number is the cost. In many documented deployments, AI infrastructure spend simply replaced the human labor it was meant to offset, a zero-sum outcome dressed up as innovation. The reason is structural: organizations running inference through hyperscaler cloud APIs at scale are paying what we call the Compute Tax,

and in 2026, the smartest operators will begin to refuse it.

The pivot to Infrastructure Sovereignty means repatriating AI workloads, moving from cloud-dependent large language models toward quantized Small Language Models running on bare metal, optimized for specific production tasks at the edge. The immediate benefit is economic: inference at the edge and CDN offloading deliver material Opex reductions. The strategic benefit is more significant: sensitive IP (scripts, dailies, unreleased footage) never leaves the organization’s controlled environment.

2026 will be, in practice, the “Year of the Sledgehammer POC”. Many organizations will still brute-force through cloud LLMs – models running at 100B to 1 trillion parameters – for initial workloads, and that is a rational short-term choice. But it is an expensive one, and the efficiency curve is moving fast. Small Language Models (SLMs), running at 1B to 8B parameters, already cover 90% of narrow M&E use cases – metadata tagging, QC triage, localization, compliance monitoring – at a fraction of the compute cost, and at 3B parameters they hit the operational sweet spot for most production tasks. Beyond SLMs, emerging architectures point further still: Hierarchical Reasoning Models (HRMs) separate abstract strategy from rapid computation, dramatically reducing overhead for complex workflows, while Tiny Recursive Models (TRMs) demonstrate that recursive self-correction can achieve comparable results with a fraction of the parameters. The trajectory is clear. Those who invest now in right-sized, NPU-optimized local inference are building a durable cost and security advantage. Those who do not are building a dependency that will become structurally expensive by 2027.

Energy efficiency belongs in this conversation too. The Measurement Mandate, driven by organizations like Greening of Streaming, is pushing

power consumption from CSR reports into core technical roadmaps. Efficient inference is no longer a sustainability talking point. It is a P&L line.

4. Max Attention: The Industry's Primary Weapon Against Churn

The myth of infinite audience expansion is over. Streaming saturation is real, documented, and no longer deniable. The question for 2026 is not how to reach ten times more viewers, it is how to retain the ones you have with ten times more precision.

Attention Optimization is the reframing the industry needs. AI becomes the primary operational tool not for acquisition but for attrition prevention, predicting churn signals before a user consciously decides to disengage and delivering hyper-relevant content injections at the exact moment of risk. This requires genuine behavioral modelling at the individual level, not segment-level approximations dressed up as personalization. The key enabler is Visual Semantic Search (VSS) – a technique that indexes content against scene-level visual and emotional embeddings rather than keyword metadata, meaning the system understands what a piece of content feels like, not just how it is tagged. Applied to churn prevention, VSS allows the recommendation engine to surface content that genuinely matches a viewer's mood and context at the precise moment of disengagement risk, with a precision that metadata-driven systems cannot approach. The operators rebuilding their engagement architectures around VSS are not optimizing a feature. They are building the primary retention system of the next decade.

5. The Conversational Layer: Whoever Controls Discovery Controls Everything

The remote control is dying. Not the physical device – the conceptual model. The idea that a viewer navigates a grid, enters search terms, and browses

“Whoever controls the conversational layer controls discovery – turning the remote control from hardware into the front line of platform power.”

categories is being dismantled by a generation of users who interact with AI conversationally across every other digital surface in their lives. The living room is the last frontier, and 2026 marks its opening.

LLMs have matured into credible Conversational Control Layers for the TV experience. Viewers express intent – “show me something like what I watched last week, but lighter” or “find the series everyone is talking about after last night's game” – and agents respond with discovery that draws on rich linked data semantics, social signals, and behavioural context simultaneously. Critically, this conversational layer does not require a trillion-parameter cloud model to function. A locally-run SLM at 3B to 8B parameters is sufficient to handle natural language intent, query a video library, and respond in real time – bringing the conversational layer inside the operator's own infrastructure, with no cloud latency and no IP exposure.

This is not a UX improvement. It is a power shift. The platform that owns the conversational layer owns the primary relationship with the viewer. Broadcasters and operators who cede this layer – to a device OS, to a smart TV manufacturer, to a third-party aggregator – become wholesale pipes. The content still flows through them, but the relationship, the data, and the long-term economic value does not.

“Whoever controls the conversational layer controls discovery – turning the remote control from hardware into the front line of platform power.”

The Mandatory Path Forward

The industry's obsession with AI features was not wrong – it was just early and poorly directed. The experimentation of 2025 generated real learning, even when it generated disappointing ROI. The difference in 2026 is that the cost of continued experimentation without structural commitment is no longer a missed opportunity. It is an existential risk.

Three realities define the mandatory path. First: those who continue to bolt AI onto legacy silos will drown in complexity and escalating Compute Tax, while those who rebuilt around autonomous workflows will finally realize the 10x ROI that has remained elusive. Second: if broadcasters and operators do not control the Intent Agent – the conversational layer through which viewers discover and access content – they will be relegated to the status of wholesale infrastructure, permanently. Third: the window to establish Infrastructure Sovereignty and adopt AI-first open protocols is 2026. Organizations that wait for 2027 will be building on someone else's economic terms.

The industry's heyday was not built by organizations that waited for the market to stabilize before committing. It was built by those who defined the next era before it arrived. That option is still available. But it is closing. **CSI**



Fernando Amendola,
The Media League,
Co-founder



Thierry Fautier,
The Media League,
Co-founder