Author Name: John Gorton
Author Title: Head of Microsoft Practice & Solutions
Date: December 2025
Version 1

THE CHANGING LANDSCAPE OF MICROSOFT SAAS BILLING: FROM PER-USER TO CONSUMPTION-BASED MODELS

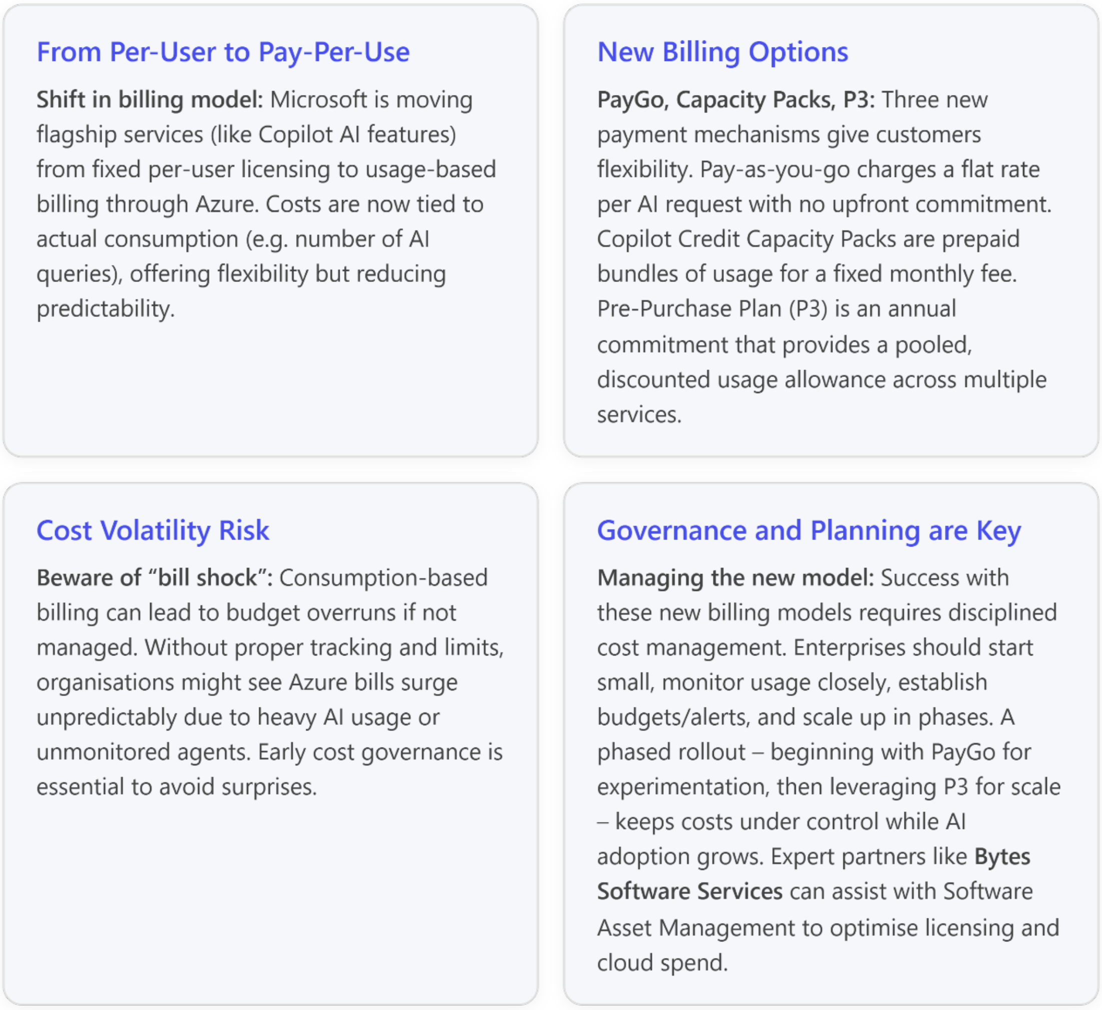
Bytes Software Services

Contents

Executive Summary

Microsoft is transforming how it bills for enterprise software-as-a-service (SaaS) offerings, shifting from predictable per-user licensing to **consumption-based billing** that runs through Azure. This evolution is most evident with AI-powered services like Microsoft 365 Copilot and custom **agents** built in Copilot Studio, where new models – **Pay-as-You-Go (PayGo)**, **Copilot Credit Capacity Packs**, and **Pre-Purchase Plans (P3)** – charge based on actual usage (e.g. number of AI "messages" processed) rather than flat per-user fees. These models offer greater flexibility and scalability, but also introduce **cost volatility**: without careful management, organisations could face unexpected "bill shock" from usage spikes or insufficient cost tracking. Looking ahead, Microsoft's recently announced **Agent 365** framework (the control plane for managing AI agents) suggests that not only users, but **AI agents themselves may require licensing** (for identity and access) – potentially adding new licensing requirements (e.g. Entra ID or Microsoft 365 licenses for agents) on top of consumption charges.

To navigate this new landscape, enterprises should adopt a **proactive cost management strategy**. This includes selecting the right billing model for each stage of adoption (e.g. start with PayGo, then consider Capacity Packs or P3 as usage grows), implementing robust tracking and controls (using Azure Cost Management, Copilot dashboards, budgets/alerts, and governance policies to prevent overruns), and optimising when to use consumption billing vs. traditional user licenses. **By taking a phased approach to rollout and cost governance, organisations can capture the benefits of AI-driven solutions while keeping expenses predictable and transparent.**



From Per-User to Pay-Per-Use

Shift in billing model: Microsoft is moving flagship services (like Copilot AI features) from fixed per-user licensing to usage-based billing through Azure. Costs are now tied to actual consumption (e.g. number of AI queries), offering flexibility but reducing predictability.

New Billing Options

PayGo, Capacity Packs, P3: Three new payment mechanisms give customers flexibility. Pay-as-you-go charges a flat rate per AI request with no upfront commitment. Copilot Credit Capacity Packs are prepaid bundles of usage for a fixed monthly fee. Pre-Purchase Plan (P3) is an annual commitment that provides a pooled, discounted usage allowance across multiple services.

Cost Volatility Risk

Beware of "bill shock": Consumption-based billing can lead to budget overruns if not managed. Without proper tracking and limits, organisations might see Azure bills surge unpredictably due to heavy AI usage or unmonitored agents. Early cost governance is essential to avoid surprises.

Governance and Planning are Key

Managing the new model: Success with these new billing models requires disciplined cost management. Enterprises should start small, monitor usage closely, establish budgets/alerts, and scale up in phases. A phased rollout – beginning with PayGo for experimentation, then leveraging P3 for scale – keeps costs under control while AI adoption grows. Expert partners like **Bytes Software Services** can assist with Software Asset Management to optimise licensing and cloud spend.

From Per-User Licensing to Azure-Based Consumption Billing

For decades, Microsoft's enterprise software licensing was **per user (per seat)** – for example, organisations would pay a fixed fee per user for Office 365 or Microsoft 365 subscriptions. This model made costs predictable, since you could forecast expenses by counting users. However, with the rise of AI-driven features and **Copilot** services, Microsoft is introducing **consumption-based billing** similar to cloud platforms. Instead of buying a license for each user that grants unlimited use, organisations now have the option to **pay based on actual usage** of certain features (measured in units like "messages" or API calls). These usage-based charges are billed through an Azure subscription (part of the company's Azure cloud billing), rather than through the traditional Microsoft licensing agreement. In other words, the Copilot/agent services blur the line between Microsoft's SaaS and cloud: your Azure bill will reflect Copilot usage costs, much like it reflects Azure VM or storage usage.

Why the change? AI services have highly variable usage patterns. One month, a team might hardly use an AI assistant; the next, they might rely on it heavily to automate tasks. A fixed per-user fee (e.g. $30/user for M365 Copilot) doesn't account for this variability – some users may overpay for little usage, while others might underpay relative to heavy usage. **Consumption-based models align cost with actual value received**: you pay more only when the service is used more. This flexibility is standard in cloud computing (AWS, Azure, Google Cloud have always charged per usage for infrastructure). Microsoft is now bringing that paradigm to its SaaS offerings to accommodate AI workloads that can scale rapidly.

However, this transition means **losing some cost predictability.** With per-user licensing, finance teams could budget a stable amount each year. Now, if usage surges (say an internal Copilot agent becomes very popular), the costs surge accordingly. Microsoft acknowledges this trade-off: their own Power Platform blog notes that while consumption-based plans "offer flexibility and agility, they can also introduce unexpected charges and budget overruns if not proactively managed". In effect, Microsoft is trading a bit of **cost certainty** for **on-demand scalability**.

Azure subscription linkage: To enable PayGo or P3 billing for Copilot/agents, an organisation must link the service to an Azure subscription and resource group in the Microsoft 365 Admin Center. All usage charges then flow into that Azure subscription's bill. This means **IT administrators and cloud finance teams need to coordinate** – costs for what was once "Office software" now appear in Azure cost management dashboards. Companies may need to adjust internal chargeback models (e.g. attributing Copilot costs to the departments that use them) and set up Azure cost alerts specifically for these new services.

In summary, Microsoft's move to Azure-based consumption billing brings it in line with cloud industry practices, enabling more fine-grained **"pay for what you use"** pricing. But it requires customers to adopt cloud-style cost management for what used to be a straightforward license count.

New Billing Mechanisms: PayGo, Capacity Packs, and P3 Explained

Microsoft has introduced three primary consumption-based billing mechanisms for Copilot and related AI agent services. Each works a bit differently:

Pay-As-You-Go (PayGo)

How it works: PayGo is a pure usage-based model with **no upfront commitment**. The organisation is charged a fixed rate per unit of consumption – in this case, per Copilot **message** (each prompt or response processed by an AI agent counts as one message). The rate is currently **US $0.01 per message**. All usage is metered and **billed through the linked Azure subscription** on a periodic basis (e.g. monthly). If no one uses the Copilot or agent in a given month, nothing is charged. If usage spikes, the costs accrue linearly with each message.

Benefits: PayGo offers maximum flexibility. It's ideal for **unpredictable or low-volume usage** scenarios. Teams can experiment with AI features without any upfront investment – you "pay $0.01 per message" and that's it. This is great for pilot projects or initial rollouts where you don't yet know how much the service will be used. It also avoids the risk of paying for capacity you don't end up using.

Trade-offs: The flip side is **cost uncertainty**. Every message has a cost, so expenses can mount if usage grows (e.g. if hundreds of employees start interacting with a Copilot agent daily). At $0.01 per message, 100,000 messages would cost $1,000 – not a huge sum in isolation, but if usage balloons further, it can impact budgets unexpectedly. There are **no built-in volume discounts** in pure PayGo; it's a flat rate. Organisations might find PayGo becomes expensive at scale (hence Microsoft provides other models for higher volumes). Additionally, PayGo requires an Azure subscription to be set up for billing (an admin must enable the PayGo meter for Copilot in the M365 admin centre and attach an Azure subscription).

In short, PayGo is **"pay for what you use, as you go"** – best for initial exploration or sporadic use because it has the lowest barrier to entry, but you must closely monitor it to avoid overspending.

Copilot Credit Capacity Packs (Prepaid Message Packs)

How it works: Capacity Packs are a **prepaid subscription for a fixed amount of usage credits**. In the case of Microsoft 365 Copilot/Studio agents, the pack is typically **25,000 Copilot credits (messages) for US $200 per month**. You purchase these "Copilot Credit" packs through the Microsoft 365 admin centre (just like buying a licensing SKU) and they are applied at the tenant level. Each pack entitles the tenant to consume up to 25k messages that month without incurring PayGo charges (the cost is covered by the prepaid fee). If you need more, you can buy multiple packs (e.g. two packs would cost $400 and cover 50k messages). Unused capacity in a pack generally **does not carry over** (it's use-it-or-lose-it each month), similar to a mobile phone data pack.

Benefits: Prepaid capacity packs make costs **more predictable** for a known baseline of usage. If you anticipate a steady volume (e.g. ~25k messages a month), buying a pack fixes your cost at $200 for that capacity, likely with some savings versus PayGo. Indeed, the effective rate per message in a $200/25k pack is $0.008 per message, which is a 20% discount on the $0.01 PayGo rate. This **discount** rewards you for the commitment. It also simplifies budgeting – e.g. always $200/month – and avoids the scenario of an unexpectedly massive Azure bill, as long as usage stays within the prepaid allotment. In Microsoft's words, the message pack is **"for more predictable usage"**, offering a tenant-level volume that you manage proactively.

Trade-offs: The organisation pays the $200 regardless of actual usage. If you overestimate needs and only use, say, 5,000 messages that month, you still pay the full amount – effectively paying $0.04 per message in that scenario, which is higher than PayGo. So, there's a risk of **overprovisioning** and wasting budget. It requires good forecasting or gradual adjustment (Microsoft allows buying additional packs as needed, so you might start with one pack and add more as usage grows). Also, capacity packs are still relatively small scale; if you need huge volume or want multi-service coverage, P3 (below) might be more suitable.

In summary, **Copilot Credit Capacity Packs** are a way to **pre-buy a chunk of usage at a discount**, ensuring a fixed monthly cost for that capacity. They are well-suited for moderate, **steady usage patterns** where you want cost stability and a slightly lower unit cost than pure pay-per-use.

Pre-Purchase Plan (P3)

How it works: The **Pre-Purchase Plan (P3)** is essentially an **annual subscription for Copilot usage credits with tiered discounts**, aimed at large-scale deployments. An organisation makes a **one-year upfront commitment** to purchase a pool of "Copilot Credit Commit Units" (often abbreviated CCCU). This pool represents a bulk quantity of messages (credits) that can be consumed across multiple services and workloads. P3 covers **Copilot Studio agents, Microsoft 365 Copilot, and Dynamics 365 Copilot** all under one committed **meter**. In exchange for the upfront commitment, Microsoft provides built-in **tiered discounts** – meaning the per-message cost is lower than both PayGo and the monthly packs, with better rates at higher volumes.

The P3 plan functions somewhat like a cloud reserved capacity: you draw down against your purchased credit pool as you use Copilot/agent services. If you consume more than the pool amount within the year, overage is billed at a pay-go rate (or you could purchase an additional P3 to top up). Notably, **P3 counts toward your Microsoft Azure Consumption Commitment (MACC)** – this is important for enterprises with an Azure committed spend deal, as investing in P3 helps fulfil that commitment (and thus may indirectly grant further financial benefits or discounts at the Azure contract level). The P3 auto-renews annually, unless adjusted, which encourages continual commitment (customers would renegotiate based on projected needs each year).

Benefits: P3 is ideal for organisations that are **ready to scale** their AI agent usage broadly and want **cost efficiency and unified management**. It offers the **lowest unit costs** via discounted bulk credits, making high volumes more economical. It also **consolidates consumption across multiple workloads** – if you have agents in Teams via Copilot Studio, plus Dynamics 365 copilots, etc., all their usage pours into one bucket. This consolidated metering can simplify cost management (one pool to monitor) and likely yields **cross-workload savings**. Microsoft describes P3 as "perfect for organizations with growing, variable demand across Copilot Studio, Dynamics 365, and M365 Copilot," providing predictable costs through a single pool. By planning an annual capacity, companies can secure better pricing and not worry month-to-month about fluctuations (as long as they stay within the commit on an annualized basis).

Trade-offs: P3 requires a **significant upfront commitment** (financially and in terms of estimating usage a year out). It is less flexible – if your AI adoption doesn't grow as fast as expected, you might under-use your credits (though you could try to push more usage to utilise what you paid for). Conversely, if adoption greatly exceeds expectations, you'll pay overage or need to buy more mid-year. In short, **forecasting risk** is the big factor. P3 is also only cost-effective if you truly have large, continuous usage across the enterprise; smaller deployments wouldn't justify it. Finally, because it auto-renews, organisations must remember to adjust or cancel if needs change – otherwise they might keep paying for a large block they no longer need.

In summary, P3 is a way to **"lock in" a year's worth of AI usage at a discounted rate**, best suited for mature deployments of Copilot/agents at scale. It shifts Copilot billing into an annual operating expense that can be budgeted with high confidence (albeit with upfront payment). Think of it as analogous to a cloud provider's 1-year or 3-year reservation – trading flexibility for a lower price and predictability.

Choosing Between the Models

Each of these billing mechanisms meets a different need on the spectrum of scalability vs. predictability. Microsoft's guidance suggests a progressive approach: *"start with PAYG … then migrate to a combination of Prepaid and PAYG as usage ramps … move to … Pre-Purchase Plan for cost savings at scale"*. In practice, organisations might use **PayGo in early phases** to get immediate access and data on usage patterns, then adopt **Capacity Packs** once they can predict a baseline of messages per month and eventually consider **P3 for enterprise-wide rollouts** where the highest discounts and centralized management matter.

It's worth noting that **traditional User Licensing (M365 Copilot license)** remains an option in parallel to these consumption models. The Microsoft 365 Copilot add-on license (currently $30 per user/month) gives a user unlimited access to Copilot features and even the ability to build agents, with **"zero-rated" agent message consumption for that licensed user**. This means if every user who uses Copilot is fully licensed, you might not need to pay per message at all – the cost is absorbed in the license. In reality, organisations will likely have a mix: you might license certain heavy users (or power builders) and use consumption billing to cover broader occasional use by others. Microsoft explicitly allows mixing and switching models – for example, you can start someone on PayGo and later assign them a license if that becomes more cost-effective. The new billing flexibility is about matching cost to actual usage patterns.

Table 1 – Microsoft Copilot Billing Models at a Glance (costs in USD):

Billing Model	**Payment & Billing**	**Price**	**Pros**	**Cons**	**Best For**
Per-User License	Per user per month (add-on license)	$30 per user/month [microsoft.com]	Predictable cost per user; unlimited usage for that user; simple to manage licensing.	Expensive if user's usage is low; not available to unlicensed users; requires upfront license purchase.	Established, heavy individual users who consistently use Copilot features (cost is justified by continuous use).
Pay-As-You-Go (PAYG)	Consumption-based via Azure meter	$0.01 per message	No upfront cost: pay exactly for what is used; very flexible for pilots or variable use; open to all users (incl. those without licenses).	Harder to predict total cost; no volume discount at high usage; requires Azure setup and monitoring to prevent overspend.	Pilot projects, small-scale trials, or unpredictable usage scenarios; extending Copilot to a wide audience for a short-term test.
Capacity Pack	Prepaid monthly subscription (tenant-level)	$200 per 25k messages per month	Cost stability month-to-month; ~20% cheaper per message than PAYG; easy to purchase via Admin Center; supports multiple users collectively.	Need to estimate usage (risk of under- or over-utilisation each month); unused messages don't roll over; still limited in scope (packs can be stacked but each is fixed size).	Departments or solutions with fairly steady monthly usage of Copilot/agents; budgeting a fixed monthly AI cost while allowing some growth headroom.
Pre-Purchase Plan (P3)	Prepaid annual commitment (Azure billing)	Custom annual purchase (discounted tiered	Lowest unit cost via tiered discounts; one pool for all	Requires large upfront spend; commitment may	Enterprise-wide AI deployments and long-term adoption;

		pricing, e.g. millions of messages/year)	workloads; counts toward Azure commitment; predictable annual cost for AI at scale.	exceed actual need; less flexibility if needs change mid-year; enterprise agreement process needed.	organisations with high volumes across many users and apps, seeking to optimise cost and centralise management.

The Impact: Cost Volatility and "Bill Shock" in a Consumption World

Shifting to consumption-based billing has significant implications for financial management. With per-user licensing, you knew your cost upfront (e.g. 100 users * £x each = £y per year). Now, cost is a function of **usage metrics** that can vary widely. This introduces the risk of **cost volatility** – month-to-month expenses can swing up or down based on user behaviour or adoption trends. A successful pilot that suddenly gains thousands of interactions could **generate a sudden spike in consumption charges**. If such usage wasn't anticipated, the result could be a budget overrun, often called "bill shock" when the invoice arrives higher than expected.

Several factors contribute to this volatility:

- **Unpredictable user adoption:** It can be hard to forecast how much employees will use a new AI tool. If Copilot truly helps productivity, usage might ramp up exponentially (each user asking more questions, integrating it into daily work). Without **usage caps or alerts**, this enthusiasm translates directly into higher costs. Microsoft's own guidance warns that PayGo and prepaid models, while flexible, "can also introduce unexpected charges, complex budgeting, and governance challenges" if not managed.

- **Immature cost-tracking for new services:** Many organisations have mature tracking for things like Azure VMs or licensed software counts, but **AI message consumption is a new metric**. Early on, IT finance teams might not have the right tools or processes to monitor "number of Copilot messages" in near-real-time. Microsoft is rolling out reporting in the admin centre (the Copilot Control System provides consumption reports and alerts), but these capabilities are new. Inconsistent or delayed metrics (e.g. if usage data updates slowly) can hamper an organisation's ability to react quickly to a surge. Companies might find their internal systems (or habits) for cost management need updating to track these new units of measure.

- **Usage variability and seasonality:** AI agent usage might fluctuate with business cycles. For example, an agent that helps with quarterly financial closing tasks may see **spikes at quarter-end** and little usage in between. If budgeting isn't attuned to these patterns, the spikes could exceed monthly allotments (incurring overage fees) whereas the quiet periods still incur base costs if using capacity packs. Such **uneven demand** makes budgeting tricky – do you size for the peak, average, or minimum? Each choice has cost implications (potentially paying for unused capacity or risking overage).

- **Multiple services and "hidden" consumption:** With Agent 365 on the horizon, organisations could soon have dozens of agents running in various departments. Each of these might consume Copilot credits under the hood. Without a unified view, costs can **add up in hidden ways**, distributed across different projects. One agent might be using far more messages due to a design issue (e.g. a loop generating many calls) – effectively a **runaway cost scenario** if not caught. This is analogous to a rogue cloud workload generating a surprise bill. It requires new governance: monitoring at the agent level and perhaps enforcing quotas (features Microsoft is looking to introduce, such as per-agent consumption limits in future updates).

All of this means that **financial governance needs to catch up** to the new billing model. The risk is very real: an unwary enterprise could roll out Copilot to thousands of employees via PayGo and only realise at month's end that millions of messages were processed, costing tens of thousands of dollars beyond budget. Early adopters have indeed raised concerns about such "bill shock" scenarios in the absence of strong cost controls.

Managing Cost Volatility: The good news is Microsoft is providing tools to mitigate these risks, and there are established cloud cost management practices that can be applied. The Microsoft 365 admin centre and Power Platform Admin Center now include **detailed consumption reports, real-time usage dashboards, and alerting**. An admin can see how many messages are being used, by whom or by which agent, and set up alerts for anomalously high usage. Additionally, Azure Cost Management can be leveraged since all PayGo/P3 charges flow into an Azure subscription – Azure's native budgeting tool allows setting spend thresholds (e.g. alert when Copilot costs exceed £X in a month) and even automated actions (though automated cost cutoff for Copilot might require custom configuration). Microsoft's Agent Cost Management eBook emphasizes **forecasting and tracking**: estimate in advance, monitor continuously, and be ready to adjust billing models if needed.

Nonetheless, these are relatively new capabilities, and organisations must actively use them. **Governance policies** will need to be established, such as: who is allowed to enable new agents (to control unplanned proliferation), who receives cost alerts, and how to respond if usage is trending above budget (e.g. deciding when to purchase extra capacity or disable a costly feature). It's also important to educate stakeholders – "pay-per-use" is a mindset shift for business units used to fixed licenses, so finance teams and department heads should be made aware of this variability. IT should consider implementing **cost guardrails**, for example: *require approval to switch on Copilot for more than X users on PayGo*, or *set a limit on messages per user per day if possible* to prevent inadvertent overuse.

In summary, consumption-based billing brings a double-edged sword: **you only pay for what you use, but you *pay* for *everything* you use.** The onus is on the customer to monitor and manage that usage. Without proper cost governance, the flexibility can indeed lead to nasty surprises. With governance, however, comes the ability to optimize costs dynamically – turning off or tuning down usage when not needed, and scaling up when the value justifies it. We will discuss strategies to manage and mitigate these risks in a later section ("Managing Costs…"), but first, another emerging consideration in this new landscape is how *licensing* itself might evolve for AI agents.

Microsoft Agent 365 and Future Licensing Implications for AI Agents

Announced at Ignite 2025, **Microsoft Agent 365** is a new management platform for AI agents – essentially a "control plane" to register, monitor, and govern all the autonomous agents running in an organisation. While its primary focus is on security and management (providing a unified registry, access control, telemetry, etc.), Agent 365 introduces the notion that **each AI agent is an entity to be managed much like a user** in your directory. In fact, Microsoft's vision is "to manage agents the way you manage people, using the same infrastructure, apps, and protections that power your business today". This may have direct licensing and cost implications:

Agent identities (Entra ID): Agent 365 requires that every agent be assigned a unique **Entra ID (Azure AD) identity** – often called an **"agent ID"**. This means when you create or register an AI agent, it gets credentials in your directory, just like a new employee or service account. Judson Althoff (Microsoft's commercial business lead) described it as *"in the same way you provision an identity for a new employee…you'll provision identity and access controls for your agents."* This approach is logical for security (it allows using Azure AD policies to control the agent's permissions, apply conditional access, etc.), but it blurs the line between human users and AI agents in licensing terms. Many Azure AD (Entra ID) features – for example, Conditional Access or Identity Protection – are only available if the entity is covered by a certain license (like Entra ID Premium P1/P2, often part of EMS or Microsoft 365 E5 suites). **Will each agent's Entra account require a paid Entra ID license?** It's a strong possibility. Microsoft has not explicitly confirmed licensing for agent identities yet (Agent 365 is in preview and "Microsoft has not finalized pricing" as of Ignite 2025), but enterprises should anticipate that agents, being first-class directory objects, might need licensing similar to any other user for certain functionality. This could mean additional Azure AD/Entra ID costs per agent, especially if you want them to adhere to advanced security policies.

Microsoft 365 access for agents: Agents often need to read or write data in Microsoft 365 services (Emails, SharePoint files, Teams chats, etc.) to perform their tasks. Under current models, if a user interacts with Copilot, that user's license covers the access (Copilot acts on the user's behalf). But if an autonomous agent is acting independently, one can argue it **might need its own M365 license** to access those services legally. For example, if you have an agent that scans SharePoint and sends emails, you may need to assign that agent a Microsoft 365 license (Exchange Online plan to send mail, SharePoint plan to read files) or otherwise treat it as a service principal with appropriate licensing. We don't have official rules yet, but Microsoft's phrasing "reliably extend your infrastructure for users to agents" hints that agents become a new class of "workload principal". The **Windows Central** coverage of Agent 365 noted that *"a tenant needs to have at least one license of Microsoft 365 Copilot to use Agent 365"* during the preview. This suggests that Microsoft is tying Agent 365's availability to the purchase of Copilot licenses. It might be a simple gating for preview, but it could also foreshadow that **agents will only be fully supported in environments that have invested in Copilot licensing** (i.e. you can't bypass the $30/user Copilot license entirely by just using PayGo agents; Microsoft wants you to have some licensed base).

Potential Agent licensing model: Microsoft could introduce a new licensing model in the future where you license the agent itself. For example, there might be an **"Agent 365 license"** that grants an agent the rights to operate within the tenant, use certain features, and perhaps includes some consumption allowance. This is speculative, but the groundwork is there: Agent 365 is analogous to an identity & management system for non-human workers. Just as you license a human for M365 or Dynamics, you might license an AI worker. This license could bundle the necessary Entra ID, compliance, and even some AI execution credits. Microsoft might require, for instance, an **Entra ID P2** for each agent (to ensure it has full Conditional Access, Identity Protection coverage) and maybe a base **Microsoft 365 F1 license** if the agent needs to access basic Office 365 data. Alternatively, they might include agent entitlements as part of existing user licenses (e.g. each M365 E5 license allows you to run X number of agents).

While final details await, companies should **plan for the scenario that deploying production AI agents will carry licensing requirements beyond just consumption costs.** In practical terms, this means when budgeting for AI, consider not only the PayGo/P3 message costs but also the identity and software access for the agent. A parallel can be drawn to service accounts today: for example, a service account that needs mailbox access often requires an Exchange Online license. Similarly, an autonomous agent performing tasks in an environment might need analogous licensing.

Cost impact: If each agent requires its own set of licenses, this introduces a more **fixed cost element** into the model (e.g. £x per agent per month, on top of variable usage). It could complicate cost calculations but also add predictability – if you know you will run 5 agents, and each requires (hypothetically) a £10 license, that's £50/month fixed. The remaining cost would be their consumption (which you might rein in through design or limits). It also creates a gating factor: an organisation might limit how many agents they deploy because each one has a direct cost even if idle. Microsoft Agent 365's purpose is to help **"track every agent being used or built, eliminating blind spots"** – from a governance perspective that includes not just security but also accountability for cost and licenses.

Action for customers: Keep an eye on Microsoft's licensing announcements for Agent 365. Engage with your Bytes account manager to understand preview licensing conditions. In the interim, apply the principle of least privilege to agents – if you can run an agent under the context of a licensed user (impersonation) that might avoid needing separate licensing for now, but that might not be sustainable long-term. Prepare your asset management and compliance processes to treat agents as a new class of "user" for tracking. This is where **Bytes Software Services** can assist, by updating your Software Asset Management (SAM) framework to include AI agents and their licensing needs, ensuring you remain compliant and cost-efficient as rules evolve.

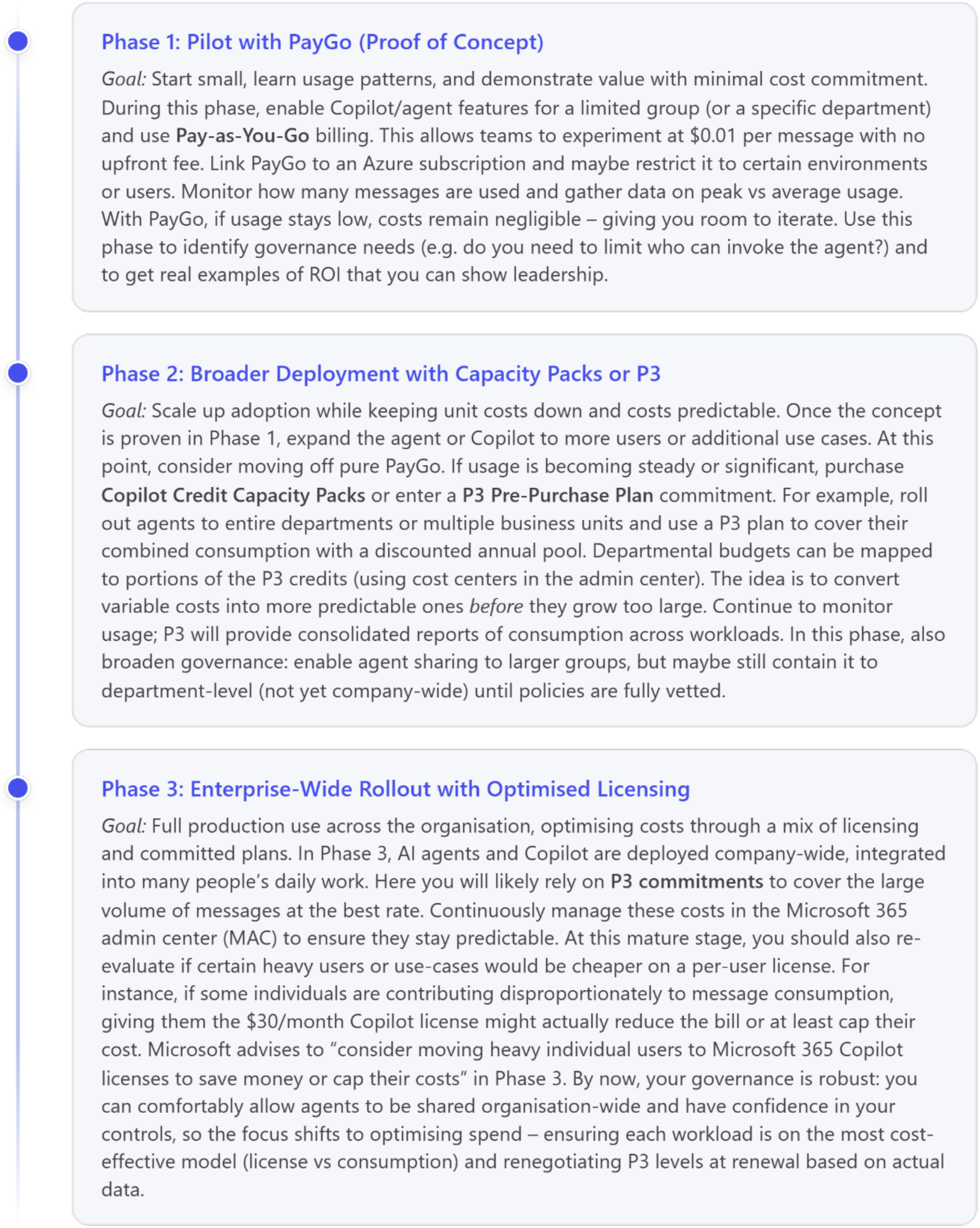
In conclusion, **Microsoft Agent 365 foreshadows a future where AI agents are first-class citizens of the IT environment – with identities, access permissions, and yes, likely licensing and costs attached.** Organisations should plan for this and include it in their cost governance strategy, so that the convenience of automated agents doesn't come with unexpected licensing compliance issues or expenses. As one industry observer quipped, "Does your AI have an ID?" – in Microsoft's ecosystem, it certainly will, and that ID won't be free.

Strategies for Managing Costs in a Consumption-Based Billing Landscape

Facing these changes, enterprise IT and finance teams need to adapt how they plan and control software costs. Below are **key strategies to manage costs** effectively under Microsoft's evolving billing models, while still encouraging innovation with AI agents and Copilots.

1. Adopt a Phased Rollout with the Right Billing Model at Each Stage

Don't try to predict a perfect end-state from day one. Instead, use a **phased approach** to deployment and billing, aligning your cost model with your stage of adoption. Microsoft itself recommends a three-phase rollout for Copilot/agents to balance learning and cost control:



Phase 1: Pilot with PayGo (Proof of Concept)

Goal: Start small, learn usage patterns, and demonstrate value with minimal cost commitment. During this phase, enable Copilot/agent features for a limited group (or a specific department) and use **Pay-as-You-Go** billing. This allows teams to experiment at $0.01 per message with no upfront fee. Link PayGo to an Azure subscription and maybe restrict it to certain environments or users. Monitor how many messages are used and gather data on peak vs average usage. With PayGo, if usage stays low, costs remain negligible – giving you room to iterate. Use this phase to identify governance needs (e.g. do you need to limit who can invoke the agent?) and to get real examples of ROI that you can show leadership.

Phase 2: Broader Deployment with Capacity Packs or P3

Goal: Scale up adoption while keeping unit costs down and costs predictable. Once the concept is proven in Phase 1, expand the agent or Copilot to more users or additional use cases. At this point, consider moving off pure PayGo. If usage is becoming steady or significant, purchase **Copilot Credit Capacity Packs** or enter a **P3 Pre-Purchase Plan** commitment. For example, roll out agents to entire departments or multiple business units and use a P3 plan to cover their combined consumption with a discounted annual pool. Departmental budgets can be mapped to portions of the P3 credits (using cost centers in the admin center). The idea is to convert variable costs into more predictable ones *before* they grow too large. Continue to monitor usage; P3 will provide consolidated reports of consumption across workloads. In this phase, also broaden governance: enable agent sharing to larger groups, but maybe still contain it to department-level (not yet company-wide) until policies are fully vetted.

Phase 3: Enterprise-Wide Rollout with Optimised Licensing

Goal: Full production use across the organisation, optimising costs through a mix of licensing and committed plans. In Phase 3, AI agents and Copilot are deployed company-wide, integrated into many people's daily work. Here you will likely rely on **P3 commitments** to cover the large volume of messages at the best rate. Continuously manage these costs in the Microsoft 365 admin center (MAC) to ensure they stay predictable. At this mature stage, you should also re-evaluate if certain heavy users or use-cases would be cheaper on a per-user license. For instance, if some individuals are contributing disproportionately to message consumption, giving them the $30/month Copilot license might actually reduce the bill or at least cap their cost. Microsoft advises to "consider moving heavy individual users to Microsoft 365 Copilot licenses to save money or cap their costs" in Phase 3. By now, your governance is robust: you can comfortably allow agents to be shared organisation-wide and have confidence in your controls, so the focus shifts to optimising spend – ensuring each workload is on the most cost-effective model (license vs consumption) and renegotiating P3 levels at renewal based on actual data.

This phased approach ensures you're not over-committing costs upfront. You **learn and adjust** as you go: start cheap and flexible, then lock in discounts as usage patterns stabilize. It provides a safety net against over-provisioning and builds executive confidence because you can show that costs are being managed stepwise (no big bang budget request without data).

2. Implement Rigorous Cost Monitoring and Alerts (FinOps for Copilot)

In a consumption model, continuous **visibility** is your best defence against surprises. Treat Copilot/agent usage with the same FinOps (Financial Operations) discipline as you would Azure or AWS cloud spend. Concretely:

- **Set up budgets and alerts in Azure Cost Management:** Since PayGo and P3 charges roll up into Azure, use Azure's budgeting tool to create a specific **Copilot/Agents cost budget**. For example, set a monthly budget threshold (maybe based on your expected usage from the pilot phase) – e.g. £1,000/month – and configure alerts at 80%, 100%, etc. Azure can email or SMS when those thresholds are crossed. This early warning gives you time to respond (investigate usage drivers, decide if it's worth exceeding the budget or needs throttling). If you have engineering resources, you can even automate responses: for instance, Azure Cost Management can trigger an Azure Function or Power Automate flow when budget is exceeded. While not natively built for Copilot yet, a creative solution could disable an optional feature or send a notification to admins to intervene if costs spike abnormally.

- **Use the Copilot usage reports (CCS and PPAC):** Microsoft's Copilot Control System and Power Platform Admin Center provide **detailed usage analytics** specifically for Copilot and agents. Ensure your admins are familiar with these dashboards. They show metrics like number of messages consumed per agent or per user, trends over time, and even allow real-time monitoring of active usage. Make it a practice to review these reports regularly (at least weekly during a rollout, and daily during initial launches or peak periods). Look for anomalies – e.g., one agent consuming far more than others, or usage times that don't align with work hours (could indicate an automated loop or misuse). Early detection of anomalies can prevent runaway costs.

- **Leverage built-in alerts and limits:** According to Microsoft's guidance, the admin centres either have or will have features like *"real-time alerts for high consumption"* and the ability to set **per-user or per-agent message caps**. If available, use these! For example, if you can set a rule that an agent cannot exceed 10,000 messages per day, that caps the financial exposure of a single rogue agent. Or if you limit each user to, say, 100 queries per day (just as a guardrail), you also bound the cost per user per day. While you don't want to stifle legitimate use, these controls can prevent accidents (like an unintended infinite loop or a user scripting the Copilot to do something repeatedly). At minimum, configure alerts: e.g. alert if any single user generates more than X messages in an hour, or if any agent's total crosses Y in a day. Future enhancements are expected in this area (Microsoft hinted at "per-agent consumption limits and budgeting controls" coming to the admin tools).

- **Tag and track by department/project:** If multiple teams or solutions are using Copilot, use **cost allocation strategies**. Microsoft allows scoping billing policies to departments or environments. Ensure each department's usage is visible to them and to management. This transparency drives accountability – if one team's usage (and hence cost) spikes, it will be evident in reports tied to their cost centre. You can even implement an internal chargeback: bill internal departments for their Copilot usage (using the data from reports) so they have skin in the game to use it efficiently. This mirrors how cloud resources are often allocated internally and can incentivize teams to stay within budget.

- **Frequent reviews and forecasts:** Make cost review a part of your AI project's routine. For instance, in your project stand-ups or weekly management meetings, include a slide on "Copilot usage vs budget this week". This keeps everyone aware. Additionally, update your **forecasts** frequently. As recommended in Microsoft's eBook, review and recalibrate your consumption estimates quarterly, especially as new features roll out that might change usage patterns. If a new Copilot capability is enabled (say, a code interpreter that might use more tokens per message), factor that in. Being proactive on forecasting means you can adjust your

P3 commitment up or down at renewal, or decide to buy extra capacity packs before you actually exceed and pay higher PayGo rates.

By treating Copilot costs with the same rigor as cloud infrastructure costs, you'll catch issues early and can optimize spend continuously. Many companies have a **FinOps team** or at least a cloud cost champion – ensure that role extends to these new SaaS consumption services. It's essentially a new line item on your Azure bill that needs the same scrutiny as any VM or database.

3. Optimise Licensing vs. Consumption Continuously

One of the trickier but high-impact strategies is to keep evaluating the **mix of licensing and consumption** to minimize total cost. Microsoft now gives multiple ways to pay; the cheapest approach might be a combination of them:

- **Identify heavy users or predictable use-cases:** As data comes in, pinpoint if there are users who consistently use Copilot far above others. For example, you might find that in a given month, 5% of users account for 50% of the messages. Those power users might be cheaper to move to a flat license. At $30/month, a user would break even with PayGo if they would otherwise generate 3,000 messages (3k * $0.01 = $30). If they're doing 10,000+ messages, the license is a bargain (and reduces load on your consumption pool). Conversely, find users with minimal usage and consider removing unneeded licenses – if someone is licensed but only asks 10 Copilot queries a month, that's not cost-effective compared to PayGo. In short, **right-size who has a license** – use them for the truly constant users and use consumption for the occasional users.

- **Match use-cases to plans:** You might use different models for different solutions. For instance, maybe your sales team's Copilot (which is heavily used daily) is funded via a P3 pool because it's high-volume and variable. But an HR chatbot used only during new hire on-boarding season might remain on PayGo since it's on-and-off. It's perfectly fine to run hybrid: Microsoft allows switching billing models on a per-environment or per-app basis. Develop a decision framework: e.g. "if expected monthly messages > X and steady, go to Pack or P3; if < X or very spiky, stay on PayGo; if user-specific intensive usage, assign license." Revisit these decisions each quarter as usage grows.

- **Utilise included capacities:** Note that some Microsoft products include Copilot capabilities without additional charge. For example, M365 Copilot Chat (the personal productivity chat) might not incur message charges for licensed users. Ensure you know which scenarios are "free" under existing licenses (sometimes known as *zero-rated* usage in Microsoft terms). If a scenario can be accomplished within a user's licensed Copilot Chat instead of a custom agent, you might prefer the licensed route. This isn't always possible if you need a custom agent but keep it in mind. Also, if you have Dynamics 365 Copilot capabilities and you're already paying for Dynamics licenses, leverage those where possible before building a custom agent from scratch that would incur new costs.

- **Watch Microsoft's bundling moves:** It's possible Microsoft will introduce bundled offers (e.g. buy X Copilot licenses, get some free message credits, or vice versa). Stay alert to any licensing program announcements. Enterprise Agreement (EA) customers might negotiate a custom deal – for instance, incorporating a certain amount of Copilot consumption into their agreement. Ensure your procurement/licensing specialists engage in those conversations, as it could significantly alter the cost equation.

The key is **flexibility** – don't set and forget your licensing choices. Make it a dynamic process to allocate licenses and consumption budgets to where they yield the best ROI.

4. Strengthen Governance and Policy Controls

Preventing cost overruns isn't just about tools – it's also about **policies and governance** that guide usage behaviour:

- **Approval processes:** Incorporate a review step before enabling Copilot or creating a new agent for a team. Have the requesting team answer questions like "How many users will use this? What value do we expect? Who will own the ongoing cost?" This doesn't have to be bureaucratic, but a simple intake can ensure awareness. It can also deter launching agents for trivial reasons. For example, an idea for an agent should be evaluated not only on technical merit but also cost/benefit – if it's going to generate 100k messages a month, is the benefit worth ~$1k monthly? This kind of approval framework keeps runaway experiments in check.

- **Training and awareness:** Educate users that while Copilot feels like an unlimited AI genie, it **incurs real costs per use**. Sometimes just making people aware is enough to modulate behaviour (e.g. they might not ask the AI the same question ten times in a row if they know it has a cost). Without stifling usage, encourage mindful use: *"treat Copilot like a resource – if you can get your answer in one well-crafted question instead of three, please do so."* In developer circles, API calls cost money, so devs are taught to code efficiently; similarly, employees can be guided to use AI efficiently. This is part of change management as AI becomes a daily tool.

- **Enforce least-privilege for agents:** This is more a security point but has cost implications. If each agent is only allowed access to the data, it truly needs, you reduce the chance it performs superfluous tasks. An over-privileged agent might wander into areas and use resources unnecessarily. Also, by quarantining or shutting down unauthorized agents (as Agent 365 will allow), you prevent unvetted tools from consuming resources. Essentially, **avoid "agent sprawl"** – a proliferation of agents that nobody owns or monitors closely. This ties back to cost: every unsanctioned agent is a potential unmonitored cost centre. Agent 365 will help inventory them so you can apply governance.

- **Periodic cost audits:** Conduct monthly or quarterly audits of Copilot usage. This could be part of a broader Cloud Governance Board meeting. The audit should ask: Are all running agents still providing value commensurate with their cost? Is any department exceeding their forecast? Did any "temporary" pilot get forgotten and left running (racking up charges)? By treating usage review as a formal exercise, you can decide to retire or pause agents that aren't ROI-positive. It's akin to turning off underused VMs to save money.

- **Leverage Bytes SAM services for governance:** Bytes Software Services (through its Software Asset Management offerings) can assist in creating the governance framework around these new consumption licenses. For instance, Bytes can help implement policies and tracking mechanisms, ensure that any new consumption service has an owner and a budget attached, and integrate Copilot billing into your existing SAM governance reviews. External experts can lend insights from other customers and best practices, ensuring you're covering all bases (e.g. they might suggest specific controls or optimisations you hadn't considered).

5. Partner with Experts and Tools for Cost Management

Managing cloud-like consumption costs might be new territory for some organisations' licensing or IT asset teams. This is where leaning on **experienced partners and advanced tools** can pay off:

- **Software Asset Management (SAM) Services:** Engage your SAM provider (like **Bytes Software Services**) to extend their oversight to Copilot and agent usage. Bytes can help by **auditing your consumption**, identifying inefficiencies, and ensuring you're in the best licensing position. For example, we can analyse your usage data and recommend an optimal mix of PayGo vs. P3 vs. licensing, based on cost projections – essentially performing a cost optimisation analysis regularly. We can also assist with forecasting models using Microsoft's Consumption Estimator tool, tailoring it with your real data to predict future spend under different scenarios.

- **Cost Management Tools:** Consider using third-party cloud cost management tools (if you already use them for Azure/AWS) to incorporate M365 Copilot costs. Many tools (Quantum, CloudHealth, etc.) allow custom cost tracking. By importing Copilot consumption as a resource, you could get combined reports of all cloud spend. Microsoft's own Azure Cost Management will eventually treat these costs like any Azure service, so you can use its dashboards and even Power BI integration for custom reports. Additionally, Microsoft is integrating Copilot billing info into their **Power Platform admin analytics** – ensure those capabilities are utilised for automated reporting (reports can be exported or scheduled to email relevant owners).

- **Continuous improvement and reviews:** In fast-evolving areas like AI billing, it helps to have periodic reviews with Microsoft or knowledgeable partners. They can update you on new features (maybe new discount offerings, or tools like the Copilot Consumption Estimator updates) and best practices. For instance, Microsoft might improve the granularity of cost data (e.g. cost per user). Knowing and using these as soon as they're available keeps you ahead of the curve.

- **Executive reporting:** When you've taken the above steps, **report up to executives on cost management wins**. Show them that, for example, *"we scaled from 100 to 1000 users of Copilot, but through proactive cost management, our cost per user only increased 20% instead of 10x"*, or that *"we identified £xm of potential annual savings by switching model/licensing"*. This builds trust and confidence at the C-level in your approach, which means they'll be more willing to greenlight further AI projects. As Microsoft's guidance notes, a proactive cost management approach "builds trust with stakeholders, making it easier to expand AI-driven solutions". Essentially, prove that cost is not spiralling out of control – it is being tamed and optimised as you grow. This reassurance can be the difference between your AI program getting expanded funding or getting put on hold due to cost fears.

Finally, recognise that cost management itself is an ongoing journey. As Microsoft introduces new billing models or changes prices (which will inevitably happen as these services mature), you'll need to adapt. By instilling a culture of **financial governance, cost transparency, and agility** in switching billing approaches, you set your organisation up to take advantage of AI advancements without overspending.

How Bytes Software Services Can Help

Navigating this new billing landscape can be challenging. **Bytes Software Services**, as a leading Microsoft partner, offers Software Asset Management (SAM) and cloud cost optimisation services that are well-suited to help customers manage and optimise Copilot and Agent 365 costs:

- *Expert Licensing Guidance*: Bytes' licensing specialists stay up to date with Microsoft's latest licensing terms and product roadmaps. They can interpret how new offerings (like Copilot, capacity packs, P3, Agent 365) apply to your specific licensing agreements. For example, if you're unsure how adding a Copilot P3 commitment fits into your Enterprise Agreement, or whether an agent identity needs a certain license, Bytes can clarify and ensure you remain compliant. We'll help craft a licensing strategy that minimises overlap and waste – perhaps by leveraging existing entitlements you might not realise you have.

- *Consumption Analysis and Optimisation*: Through SAM, Bytes can collect detailed data on your Copilot consumption and provide **analysis reports**. They might highlight, for instance, that 30% of your Copilot messages are coming from a process that could be optimized or a user group that could be shifted to a different plan. With tools and expertise, they can model scenarios (what-if analyses: "What if you purchased one more capacity pack vs. pay-go over next 6 months?") and recommend the most cost-effective approach. This kind of data-driven advice ensures you're not leaving savings on the table.

- *Cost Management Process Integration*: If your organisation doesn't have a cloud cost management practice, Bytes can help implement one. We can set up the Azure Cost Management budgets and alerts for you, configure dashboards, and even manage them on your behalf as part of a managed service. Our teams can receive the alerts and act as an early warning system, notifying you and providing remediation steps when consumption anomalies occur. Essentially, we can augment your team to handle the day-to-day monitoring of these new workloads' costs.

- *Governance and Policy Planning*: Bytes SAM consultants have seen how other organisations handle cloud and SaaS governance. We can assist in updating your IT policies to incorporate AI agent governance. This could include templates for an "AI Usage Policy", guidelines for departments requesting Copilot features, and frameworks for internal chargeback. Leveraging our experience can accelerate your governance maturity, so you're not starting from scratch. We can also facilitate stakeholder workshops to educate your teams on the new billing concepts (ensuring that both IT and Finance understand how PayGo or P3 works, for instance).

- *Value Realisation and ROI Tracking*: Part of controlling costs is also maximising the value you get from what you spend. Bytes can help you connect the dots between cost and benefit. By establishing metrics for productivity gains or time savings from Copilot, we help you demonstrate ROI. This is crucial – if you can show that a $1,000 Azure Copilot bill yielded $50,000 worth of labour hours saved, it puts costs in perspective. Bytes can assist in setting up this value tracking (perhaps through Power BI reports that combine usage data with efficiency metrics). A well-run AI project isn't just about cutting cost; it's about **optimising cost for maximum return**. Bytes outside perspective can ensure you're measuring that.

In summary, **Bytes Software Services acts as a trusted advisor and extended team** to navigate the financial side of this technological shift. With their SAM and cost management expertise, you get the dual benefit of **controlling risks (avoiding surprise costs, staying compliant)** and **optimising spend (choosing the best plans,**

proving ROI). This allows your organisation to focus on deploying AI agents to transform the business, while Bytes helps safeguard your budget and licensing compliance in the background.

Conclusion: Embracing the New Model with Eyes Wide Open

The landscape of Microsoft SaaS billing is undeniably changing – moving from the comfort of per-user licenses to the dynamism of consumption-based pricing. This shift brings exciting opportunities: **fine-grained cost alignment with usage, the ability to start small and scale fast, and potential savings through efficient use of resources.** It also brings new responsibilities: organisations must **actively manage and govern their usage** to avoid overspending and be prepared for new licensing wrinkles (like licensing AI agents) as Microsoft's ecosystem evolves.

By understanding how **PayGo, Capacity Packs, and P3 plans** work, enterprise customers can choose the right tool for the job – leveraging PayGo's agility, the predictability of prepaid packs, or the economies of scale in P3 when appropriate. Importantly, moving these services into Azure's billing domain means **traditional software asset management merges with cloud cost management**. Finance and IT teams in Microsoft shops will need to adopt the best practices that cloud-centric teams have honed: things like setting budgets, monitoring in real-time, and continually optimising cost vs. benefit. The good news is Microsoft is providing the necessary tools and guidance (from the Copilot Cost Estimator to admin centre controls) – but tools only help if they are used.

Looking ahead, as AI agents become part of the digital workforce, organisations should anticipate a world where **each AI agent might carry its own cost signature** – some in consumption of Azure credits, some in required licenses for identity or functionality. Planning for that now, in pilot stages, will prevent painful adjustments later. It's a journey many are just beginning: early adopters are learning lessons about cost spikes, and Microsoft is rapidly updating features in response. By staying informed (e.g. keeping up with Microsoft's announcements and resources like the Agent Cost Management eBook), and by partnering with experts like Bytes to fill gaps in expertise, you can navigate this journey smoothly.

In essence, **the era of "set it and forget it" licensing is ending** for Microsoft's most advanced services. But with the right approach, this new era can be even better. You gain flexibility – the ability to tie costs directly to usage and value – and you can potentially save money by not over-licensing upfront. The trade-off is diligence and governance. As our discussion has outlined, there are concrete steps to take: phased deployments, vigilant monitoring, smart use of licenses, strong policies, and leveraging SAM services. By taking these steps, you **turn a potential risk into an opportunity** – the opportunity to accelerate AI adoption confidently, knowing that costs are under control and transparent.

Now is the time for enterprise leaders to proactively engage with these new models. Those who do will find that they can **harness the power of AI copilots and agents at scale without breaking their budgets**, while those who ignore the change may face hard questions when the bills arrive. In short, embrace the change, arm yourself with data and strategy, and treat cost management as an integral part of your AI rollout. If you do, you'll not only avoid bill shock – you'll build a foundation of trust and financial governance that lets you unlock AI's full potential across your organisation.

Resources

This whitepaper drew on several Microsoft resources (e.g. official blogs and documentation) to ensure accuracy in describing the new billing mechanisms and best practices. For further reading, consider Microsoft's *Agent Cost Management eBook* for an in-depth guide, the *Copilot Credit P3* announcement for details on the pre-purchase plan, and Microsoft's documentation on setting up **Pay-as-you-go for Microsoft 365 Copilot** and **managing Copilot costs** on the Power Platform admin centre. These can provide additional insight and the latest feature updates as you plan your next steps in this new billing paradigm.

- Read the full Agent Cost Management eBook
- Try the Copilot Studio Agent Consumption Estimator
- Read about the Copilot Credit P3 best practices
- Watch the Agent Cost Controls webinar
- Listen to the Cost Controls podcast episode
- Explore cost management documentation
- Learn about Microsoft 365 Copilot Pay-As-You-Go