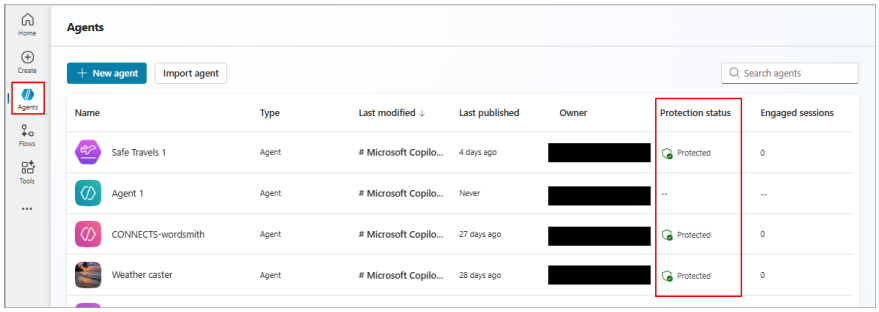
Agent runtime protection status

On the Agents page in Copilot Studio, creators can determine at the agent level whether published agents comply with security and governance standards, and verify that the agent's threat detection features are working.

For more information, see Key Concepts - Security and Administration of Copilot Studio .

The security status of published agents is displayed in the **Protection status** column on the **Agents** page . By displaying the protection status for each agent as one of the values **??Protected** , **Needs review** , or **Unknown** , creators can quickly determine when threat detection is active and which of their agents are protected.



The possible values ??for an agent published in **Protection status** are **Protected** , **Needs review** , and **Unknown** .

When an agent's agent level status is **Protected** , no immediate action is required (based on detected signals). When the status is **Protected** , a green shield icon will be displayed to visually emphasize that the agent is protected (green shield protection icon).

When the agent level status is **"Needs review** ," the agent's policy has been violated or validation is incomplete. Regardless of the status, the creator can view the agent level summary dialog details, which provide more information about the agent's security, categorized into different sections.

Select the agent status in the **Protection status** column to view the agent level summary dialog box showing the agent's protection status.

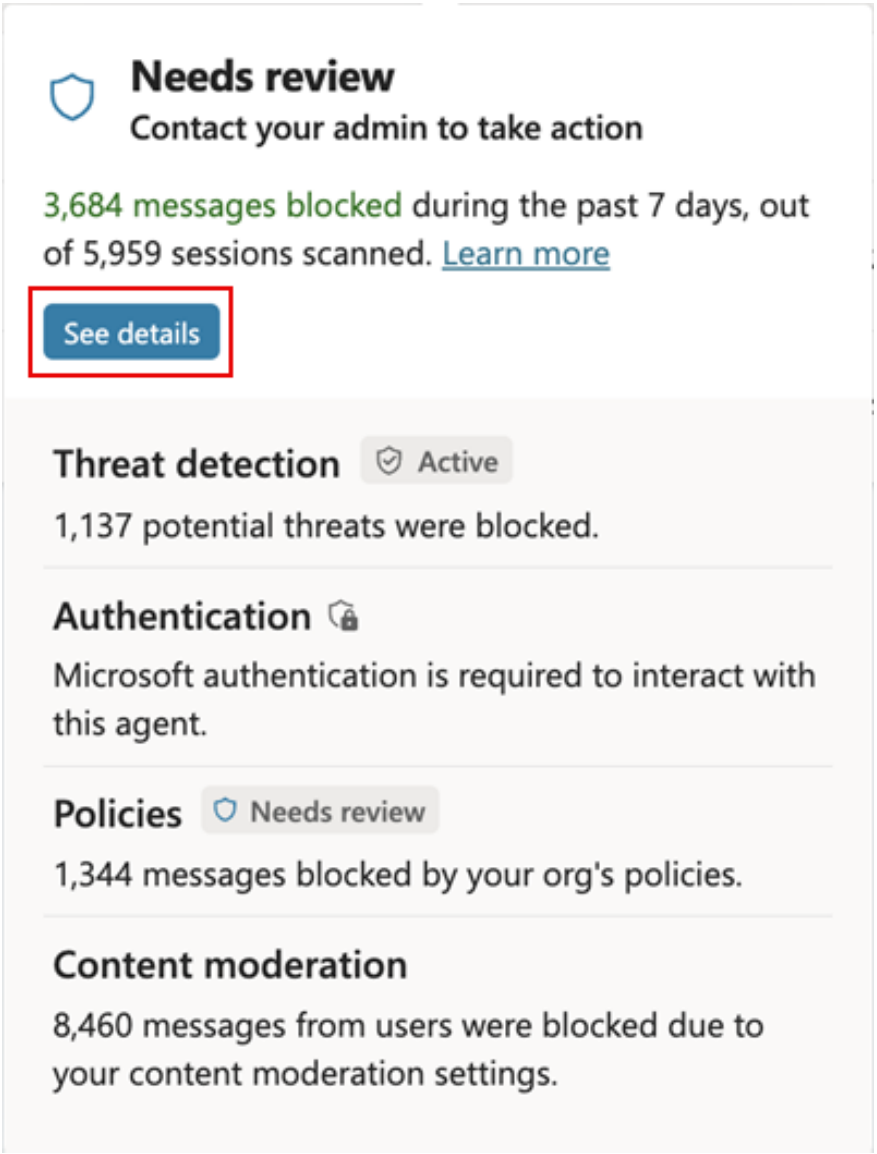
Summary of protection status

In the protection status summary dialog box, the agent's protection profile is divided into three categories: *Authentication* , *Policies* , and *Content Moderation* . Each category can have a status of **Protected** , **Needs Review** , or **Unknown** , similar to the agent-level summary status on the **Agents** page .

Additionally, this dialog box displays the number of messages blocked due to potential threats, policy violations, and content moderation settings breaches. Potential threats can be direct or indirect. For example, a potential

threat might be caused by a user overriding security settings with administrator privileges. Or, a potential threat might be a referenced knowledge source. All published agents automatically have threat detection enabled and display the **Active** label .

Similar to agent-level protection status, if **Authentication** or **Policies** violate accepted security standards, as defined by the agent's security settings, the " **Needs review"** label will appear next to the category in the summary dialog. If either of these categories has the **"Needs review"** label , this label will be aggregated into the single agent-level protection status on the **Agents** page .



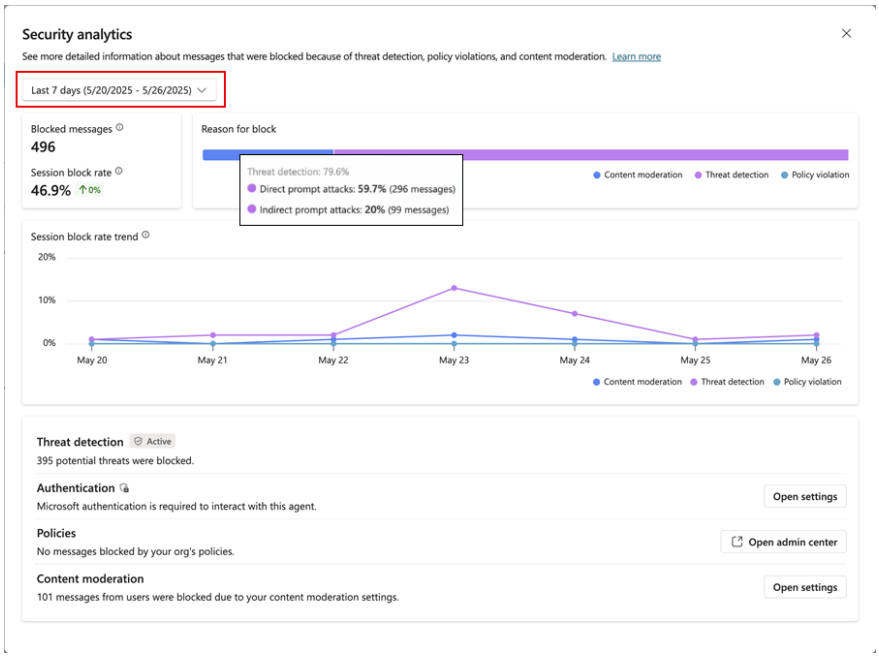
From the agent-level protection summary dialog box, if you want to see more details, select **See details to open the Security analytics** dialog box .

Security analysis

The Security Analytics dialog box displays statistics and trends on blocked messages and agent status across authentication, policy, and content moderation categories.

Select the date range of the security analysis data you want to view. If you don't have data in one or more of the following ranges, those ranges won't appear as selectable options in the date range selector. Choose from the following range options:

1. **Last 7 days**
2. **Last 14 days**
3. **Last 30 days**



The stacked Reason for Block bar chart analyzes the total number of all blocked messages within a configured timeframe into color-coded bar segments, with the length of each segment representing the percentage for each blocking reason. For example, in the chart, there are 496 blocked messages in the last 7 days. Of these blocked messages, 79.6% were blocked because they were considered potential threats (pink bar segment), with approximately 60% of all blocked messages being direct prompt attacks. A quick look at the session blocking rate trend chart shows that the rate of message blocking due to the detection of potential threats peaked on May 23rd.

The Session block rate trend chart shows the percentage of total sessions in which a prompt was blocked, as a trend line over time. Each of the three types of protection is shown individually as color-coded trend lines, using the same color coding as the **Reason for block** bar chart .

Threat detection or protection category	Describe	Act
Threat detection	This component displays statistics on blocked prompt attacks, which are proactively blocked by default. It provides manufacturers with insights into the number and trends of attacks, helping them better understand the security landscape for their agents.	

Threat detection or protection category	Describe	Act
Authentication	This component indicates whether the agent requires end-user authentication or whether it operates publicly. If an automated agent operates publicly, it could expose sensitive data to potential attackers or unauthorized individuals. In such cases, this status reflects a potential threat vector that may require the manufacturer's attention.	Select **Open settings** to go to the agent's **Settings** > **Security** > **Authentication** page , where you can choose your authentication method.
Policy	This component reflects policy violations established by administrators in the Power Platform admin center. Agents may violate these policies, and creators need to be aware of these violations and make necessary adjustments. For example, an organizational data policy might block a connector that an agent attempts to use.	Make the necessary changes to your agent to ensure it complies with your organization's policies. To review policy-related errors, select the **Review errors** in **Policies** link . If you need to change your organization's policies and you have access, select **Open admin center** to access the Power Platform admin center, where you can view or edit your organization's policies.
Content censorship	This component doesn't directly affect the protection status but is part of the statistics and trends that the producer can use. It helps ensure that the content generated by the agent adheres to the desired sensitivity levels.	Select **Open settings** to go to the agent's **Settings** > **Generative AI** page , and in **Moderation** > **Content moderation level** , adjust the slider to your desired content moderation level.