1. Q-Learning :
    - Pada setiap langkah, agen memilih tindakan berdasarkan kebijakan yang mengeksplorasi lingkungan, seperti dengan *epsilon-greedy*.
    - Setelah tindakan dipilih dan dijalankan, agen menerima *reward* dan berpindah ke keadaan baru.
    - Agen kemudian memperbarui *Q-table* berdasarkan *reward* yang diterima dan nilai Q maksimum dari keadaan baru (*greedy*).
    - Tidak seperti SARSA, Q-Learning adalah *off-policy,* artinya adalah pembaruan *Q-table* oleh agen diasumsikan bahwa agen akan selalu memilih tindakan terbaik (optimal) di masa depan.

    SARSA:

    - Agen memilih tindakan berdasarkan kebijakan saat ini.
    - Setelah tindakan dijalankan dan agen menerima *reward*, agen memilih tindakan berikutnya di keadaan baru (berdasarkan kebijakan yang sama).
    - *Q-table* kemudian diperbarui menggunakan nilai Q dari tindakan berikutnya yang telah dipilih.
    - SARSA adalah *on-policy*, artinya adalah pembaruan *Q-table* dipengaruhi oleh tindakan yang sebenarnya diambil oleh agen.

    Sederhannya, Q-Learning cenderung menghasilkan kebijakan yang lebih agresif dalam memaksimalkan *reward* karena fokusnya pada nilai maksimum masa depan dari *Q-table*, meskipun tindakan yang diambil mungkin tidak sepenuhnya mengikuti kebijakan saat ini (policy), sedangkan SARSA lebih hati-hati dan konsisten dengan kebijakan saat ini sehingga seringkali lebih lambat dalam mencapai reward maksimal. Namun, pembaruan *Q- table*-nya jauh lebih stabil.

2. Perbedaanya ada pada nilai *Q-table*-nya, pada Q-Learning, nilai Q-nya cenderung lebih tinggi daripada SARSA. Ini menunjukkan bahwa Q-Learning cenderung lebih optimis dan memilih tindakan dengan estimasi *reward* tertinggi, sedangkan Q di SARSA cenderung lebih rendah, hal ini mencerminkan pendekatan yang lebih konservatif dalam memperbarui nilai Q karena SARSA mempertimbangkan risiko atau *reward* yang akan datang dari tindakan saat ini. Selain itu, jalur yang diambil Q-Learning juga lebih panajng dibandingkan SARSA karena kecenderungan dari Q-Learning yang lebih agresif dalam mencoba jalur berbeda untuk mencari *reward* yang optimal.