MAD (mutually assured destruction)

- If player 1 does nothing, then every body lives happily ever after.
- If player 1 destroys the country of player 2, then player 2 has a threat to retaliate.
- However, no soldier has any incentive to do so after his country is destroyed.

"destroy" by player 2 is a non-credible threat.

⇒ Make "destroy" an automati response.

Movie: Dr. Strange love
<table>
<thead>
<tr>
<th></th>
<th>Cancel</th>
<th>Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work</td>
<td>10, 10</td>
<td>10, 10</td>
</tr>
<tr>
<td>Destroy</td>
<td>0, 0</td>
<td>0, -1</td>
</tr>
</tbody>
</table>
Let \( \Gamma = \langle X, A, H, P, (u_i) \rangle \) be an F6WP\( \Gamma \).

**Definition (Subgame)**

The **subgame** of \( \Gamma \) rooted at history \( h \) is the F6WP\( \Gamma (h) = \langle X, A, H|_h, P|_h, (u_i|_h) \rangle \) where

\[
H|_h = \{ h' \mid (h, h') \in H \}
\]

\[
P|_h (h') = P((h, h'))
\]

\[
u_i|_h (h') = u_i((h, h')) \quad \text{for all} \quad (h, h') \in Z
\]
Strategies relativized to histories.

For each strategy $s_i$ in $T$, let $s_i/h' := s_i((h, h'))$.

The outcome function of $T(h)$ is denoted by $D_h$. 
Def (SPE)

A subgame-perfect equilibrium (SPE) of a game \( \Gamma \) is a strategy profile \( s^* = (s_i^*) \), such that for each history \( h \in H \):

\[ s^*_{i|H} := (s_i^*|H)_{i \in N} \]

is a NE of \( \Gamma(h) \).
\[ S = (A, R) \]
\[ S = (\{ \emptyset \rightarrow A \}^3, \{ (A) \rightarrow R \}^3) \]
\[ S_{/A} = (\{ \emptyset \rightarrow A \}^3_{/A}, \{ (A) \rightarrow R \}^3_{/A}) \]
\[ H = \{ \emptyset, (A), (B), (A, L), (A, R) \} \]
\[ H_{/A} = \{ \emptyset, (L), (R) \} \]

\[ \Rightarrow S = (A, R) \]
\[ h = \emptyset, \text{s is a NE} \]
\[ h = (A), S_{/A} \text{ is a NE} \]
\[ \Rightarrow S \text{ is a SPE} \]

\[ S = (B, L) \]
\[ h = \emptyset, \text{s is a NE} \]
\[ h = (A), S \text{ is not a NE} \]
\[ \Rightarrow S \text{ is not a SPE} \]
**Example (Shelling game)**

Actions for player 1: (2:0), (1:1), (0:2)

Actions for player 2: y - agree to split
n - nobody gets anything

<table>
<thead>
<tr>
<th>Player 1</th>
<th>Player 2</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>(2:0)</td>
<td>y</td>
<td>(2:0)</td>
</tr>
<tr>
<td>(0:0)</td>
<td>y</td>
<td>(0:0)</td>
</tr>
<tr>
<td>(2:0)</td>
<td>n</td>
<td>(0:0)</td>
</tr>
<tr>
<td>(0:0)</td>
<td>y</td>
<td>(0:0)</td>
</tr>
<tr>
<td>(0:0)</td>
<td>n</td>
<td>(0:0)</td>
</tr>
</tbody>
</table>

**Table Notes:**
- **SPE** indicates a Nash equilibrium.
- **X** indicates an agreement or split.

**Diagram:**
- Tree structure with nodes for each possible action leading to outcomes for both players.

**Players’ Strategy:**
- Player 1 chooses between (2:0), (1:1), (0:2).
- Player 2 chooses between y (agree), n (split).

**Equilibrium Analysis:**
- The Nash equilibrium is marked by an `X` in the table, indicating the optimal outcomes for both players.

**Discussion Points:**
- The game highlights the strategic interplay between agreeing to split and the potential for deadlock.
- The diagram visually represents the decision tree, aiding in understanding the game's structure and outcomes.
Questions

- Does an SPE always exist?
- Under which conditions?
- How to compute it?
- What is the complexity?

We show:

- It is easy to verify that a profile is an SPE

  \[ \Rightarrow \quad \text{"one deviation property" (for finite horizon games)} \]

- For finite games, we can easily compute the SPE by "backward inclusion" (Kuhn's Theorem)
Notation: If $T$ is an E6WPS then $L(T)$ denotes the length of the longest history in $T$.

Lemma (one deviation property)

Let $T = < X, A, H, P, (u_i) >$ be a finite horizon E6WPS. Then a strategy profile $s^*$ is an SPE of $T$ if and only if for every player $i \in X$ and every history $h \in H$ for which $P(h) = i$, we have:

$$u_i(h (O_h (s^*|_h, s^*|_h)), s^*|_h) \geq u_i(h (O_h (s_i|_h, s_i)), s_i)$$

for every strategy $s_i$ of player $i$ in the subgame $T(h)$ that differs from $s_i|_h$ only in the action after the initial history of $T(h)$.

(without the underlined parts, it is just the def of SPE)
Red denotes a strategy profile.

--- denotes the branches we have to check in order to verify that we are SPI.
Proof

⇒: obvious.

⇐: By contradiction.

Suppose that $s^*$ is not an SPE. Then there is a history $h$ and a player $i$ such that $s_i$ is a profitable deviation for player $i$ in the subgame $T(h)$.

We have the number of histories $h'$ with $s_i(h') \neq s_i^{s^*}(h')$ is at most $\left\lfloor \frac{E(T(h))}{L} \right\rfloor$ and hence finite (finite horizon assumption?). Since deviations not on the result outcome paths are irrelevant.
Choose profitable deviation $s_i$ in $\Gamma(h)$ with minimal number of deviation points. Let $h^*$ be longest history in $\Gamma(h)$ with $s_i(h^*) \neq s_i(h^*)$; i.e. "deepest" deviation point for $s_i$.

Then in $\Gamma((h, h^*))$, $s_i(h^*)$ differs from $s_i(h, h^*)$ only in the initial history. Moreover, $s_i(h^*)$ is a profitable deviation in $\Gamma(h, h^*)$. Since $h^*$ is the longest history in $\Gamma(h)$ with $s_i(h^*) = s_i(h, h^*)$.

So $\Gamma((h, h^*))$ is the desired subtree where one-step deviation is sufficient to improve the utility.
The corresponding proposition for infinite horizon games does not hold.

Counter-example

\[
\begin{array}{cccccc}
\end{array}
\]

\[
\begin{array}{cccccc}
D & D & D & D & D & \cdots
\end{array}
\]

\[
\begin{array}{cccccc}
0 & 0 & 0 & 0 & 0 & \cdots
\end{array}
\]

Strategy \( s \) with \( s_i(h) = D \) for all \( h \in H \setminus \mathbb{Z} \)
- satisfies "one deviation property," but
- is not an SPE, since it is dominated by \( s^* \) with \( s^*_i(h) = A \) for all \( h \in H \setminus \mathbb{Z} \).