In recent years, deep networks have demonstrated their powerful capabilities, often outperforming human experts, such as AlphaGo. Many have been deployed to real applications, creating great business value. Because of their great impact to human beings and our society, many have concerned about their potential risks. In this talk, the speaker will first give some background information about adversarial attacks, including black-box and white-box attacks and attribution methods, including integrated gradients. Then, he will discuss how to use attribution methods, in particular integrated gradients, which was originally developed for explaining deep methods, to study risks of transferable attacks. Similar to deep network output, attributions can be manipulated by adversarial attacks. In the second half of this talk, the speaker will discuss how to protect attributions against adversarial attacks.