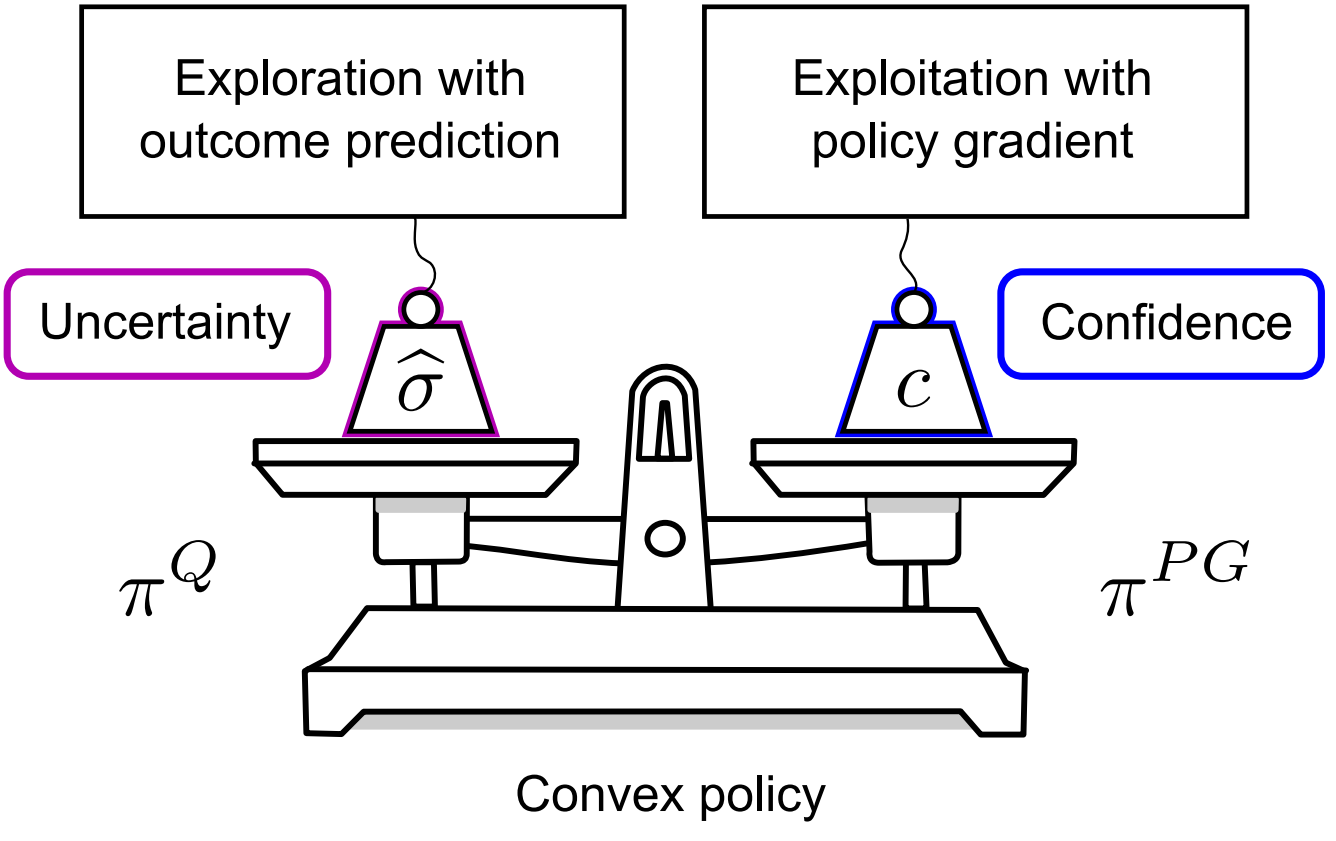
a


Exploration with outcome prediction
Exploitation with policy gradient
Uncertainty
Confidence
$\widehat{\sigma}$
c
π^{Q}
π^{PG}
Convex policy