Chris Summerfield

Training neural networks to control themselves

Control signals allow biological brains to regulate their own behaviour in the face of an uncertain environment. Despite the significance of theories of control for understanding the functioning of biological brains, the notion that brains might overtly regulate their own behaviour has been missing from theories of perception and cognition based on deep learning. I will describe an experiment in which deep networks are trained using reinforcement learning (RL) to perform a task under varying levels of uncertainty about the efficacy of their actions. I will show that training the network to predict how controllable the environment is allows them to learn to adapt optimally and on the fly to uncertain circumstances. Misinforming the network about how controllable the environment is leads to pathological behaviours in the network that resemble those observed in psychological disorders such as depression and OCD. This work offers a new way to think about healthy and disordered control in biological and artificial networks.