doi:10.1038/nature14236

Supplementary Discussion.

Average score-per-episode and predicted Q-values across learning.

To track the progress of our agent, we used two metrics: Figures 2a and 2b show the average score obtained by our agent as it evolves during learning on two games (Space Invaders and Seaquest). While the scores generally improve over time and eventually surpass human scores, as expected its progress is not perfectly steady – reflecting the noise-prone nature of the average score-per-episode metric which arises due to the strong influence of even small changes in the weights governing the policy on the overall distribution of states visited. Another more stable metric is the agent's estimated action-value function Q, which provides an estimate of how much discounted reward the agent can obtain by following its policy from any given state. Figures 2c and 2d show how the average Q-values evolve for a fixed set of game states for the same two games. The average predicted Q increases much more smoothly than the average score obtained by the agent. Note that, however, whilst the average predicted Q metric does not necessarily reflect reality, its gradual increase provides evidence of the stability and smoothness of learning in the network. This smoothness was apparent across all games and we did not observe any divergence or instability issues. Future work, however, will be necessary to determine whether theoretical guarantees of convergence exist.

Visualization of Representations Learned by DQN using t-SNE.

We also asked whether the representations learned by DQN are able to generalize to data generated from policies other than its own. To do this, we presented as input to the network game states experienced during human and agent play, recorded the representations from the last hidden layer, and visualized the embeddings generated

by the t-SNE algorithm (Extended Data Figure 1). The fact that there is similar structure in the two-dimensional embeddings corresponding to the DQN representation of states experienced during human play (Extended Data Figure 1: orange points) and DQN play (Extended Data Figure 1: blue points) suggests that the representations learned by DQN do indeed generalize to data generated from policies other than its own. Further, the presence in the t-SNE embedding of overlapping clusters of points corresponding to the network representation of states experienced during human and agent play shows that the DQN agent also follows sequences of states similar to those found in human play. Extended Data Figure 2 provides an additional illustration of how the representations learned by DQN allow it to accurately predict state and action values.

Note on Comparison of DQN Performance with Human Performance.

It is worth bearing in mind that our principal motivation for examining the performance of a professional human games tester was to provide a *reference level* against which the performance of DQN could be compared. With this aim in mind, testing of the human player was conducted under conditions that were designed to be as closely equated as possible to that of the DQN agent (see Methods). Nevertheless, it is clear that humans and artificial agents each possess different information processing strengths/weaknesses that bear on their performance on any given task, which makes it difficult to conduct a comparison that is perfectly matched across all potential factors (e.g. humans are privileged in terms of prior knowledge accrued from playing video games in the past).

Importance of Replay Memory, Target Q-Network and Deep Convolutional Network Architecture to DQN Performance.

In order to demonstrate the importance of using experience replay and a separate target network we evaluated how our system performed without these components on a subset of 5 games. The results (see Extended Data Table 3) showed that when neither experience replay nor a separate target Q-network were present in our system, the scores dropped drastically to less than 30% of the performance of the full system (i.e. with both replay and target Q-network) and even close to random agent performance in some cases. It can also be seen that using replay memory has a bigger impact on the final performance than the target Q-network, however only the experiments that used both experience replay and a separate target Q-network exhibited consistently stable learning across all games. Further, additional simulations also demonstrate the critical dependence of successful performance on the use of a deep convolutional network: when the convolutional network was substituted with a single linear layer used in combination with replay and a separate target network, performance was very poor (i.e. around chance levels: see Extended Data Table 4).