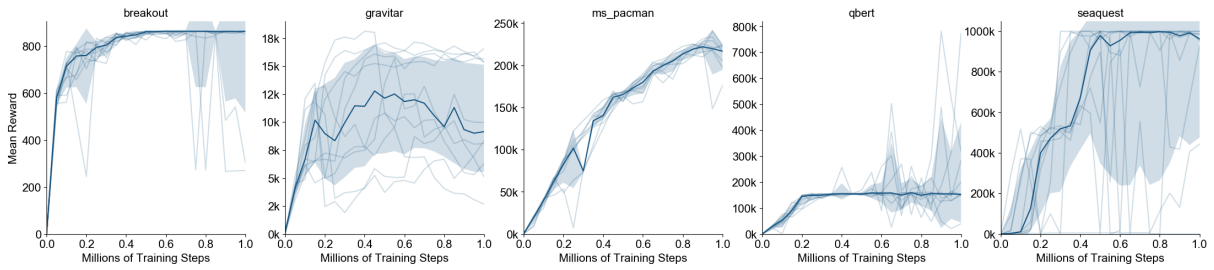

Supplementary information

Mastering Atari, Go, chess and shogi by planning with a learned model

In the format provided by the
authors and unedited



	breakout	gravitar	ms pacman	qbert	seaquest
μ	747.2	10,338.8	210,060.0	237,393.8	806,233.8
σ	230.2	4,790.4	15,520.8	193,266.1	330,129.0

Supplementary Data Figure S1: **Repeatability of *MuZero* in Atari for five games.** Total reward is shown on the y-axis, millions of training steps on the x-axis. Dark line indicates median score across 10 separate training runs, light lines indicate individual training runs, and the shaded region indicates the standard deviation across the runs. Table underneath the plot shows mean μ and standard deviation σ of total reward across the 10 training runs at the end of training.

Model

$$\left. \begin{aligned} s^0 &= h_\theta(o_1, \dots, o_t) \\ r^k, s^k &= g_\theta(s^{k-1}, a^k) \\ p^k, v^k &= f_\theta(s^k) \end{aligned} \right\} p^k, v^k, r^k = \mu_\theta(o_1, \dots, o_t, a^1, \dots, a^k)$$

Search

$$\begin{aligned} \nu_t, \pi_t &= MCTS(s_t^0, \mu_\theta) \\ a_t &\sim \pi_t \end{aligned}$$

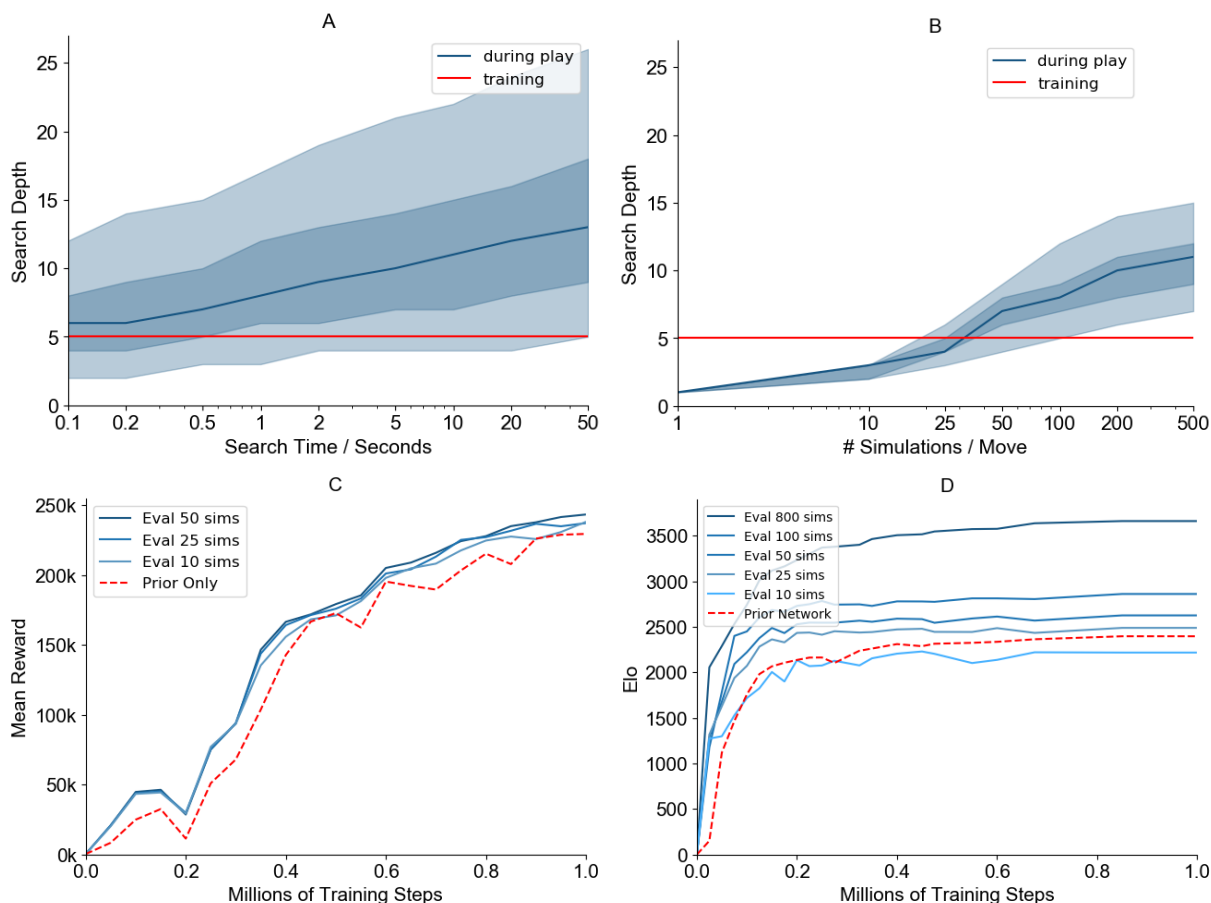
Learning Rule

$$\begin{aligned} p_t^k, v_t^k, r_t^k &= \mu_\theta(o_1, \dots, o_t, a_{t+1}, \dots, a_{t+k}) \\ z_t &= \begin{cases} u_T & \text{for games} \\ u_{t+1} + \gamma u_{t+2} + \dots + \gamma^{n-1} u_{t+n} + \gamma^n \nu_{t+n} & \text{for general MDPs} \end{cases} \\ l_t(\theta) &= \sum_{k=0}^K l^p(\pi_{t+k}, p_t^k) + \sum_{k=0}^K l^v(z_{t+k}, v_t^k) + \sum_{k=1}^K l^r(u_{t+k}, r_t^k) + c||\theta||^2 \end{aligned}$$

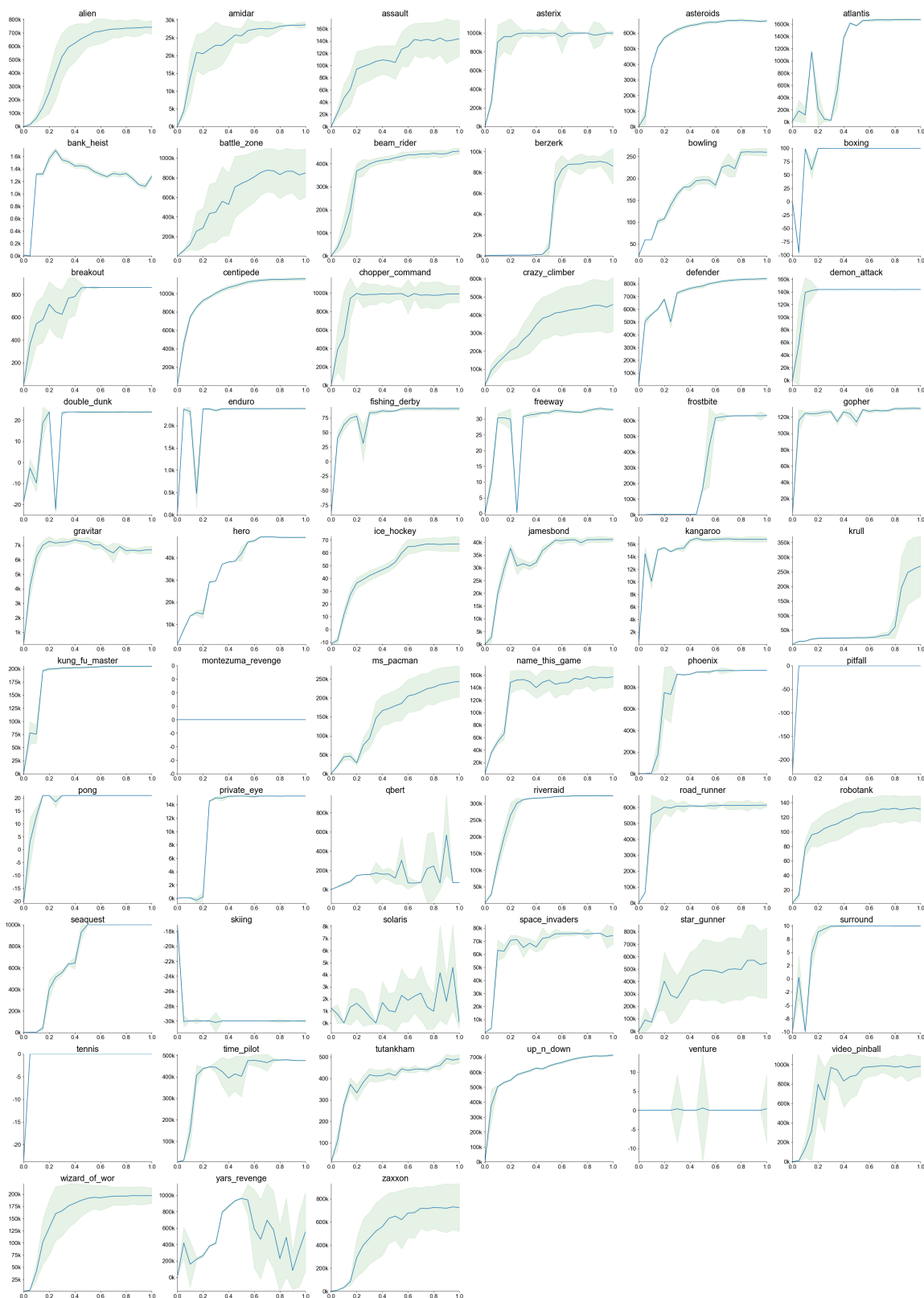
Losses

$$\begin{aligned} l^p(\pi, p) &= \pi^T \log p \\ l^v(z, v) &= \begin{cases} (z - v)^2 & \text{for games} \\ \phi(z)^T \log v & \text{for general MDPs} \end{cases} \\ l^r(u, r) &= \begin{cases} 0 & \text{for games} \\ \phi(u)^T \log r & \text{for general MDPs} \end{cases} \end{aligned}$$

Supplementary Data Figure S2: **Equations summarising the *MuZero* algorithm.** Here, $\phi(x)$ refers to the representation of a real number x through a linear combination of its adjacent integers, as described in the Network Architecture section.

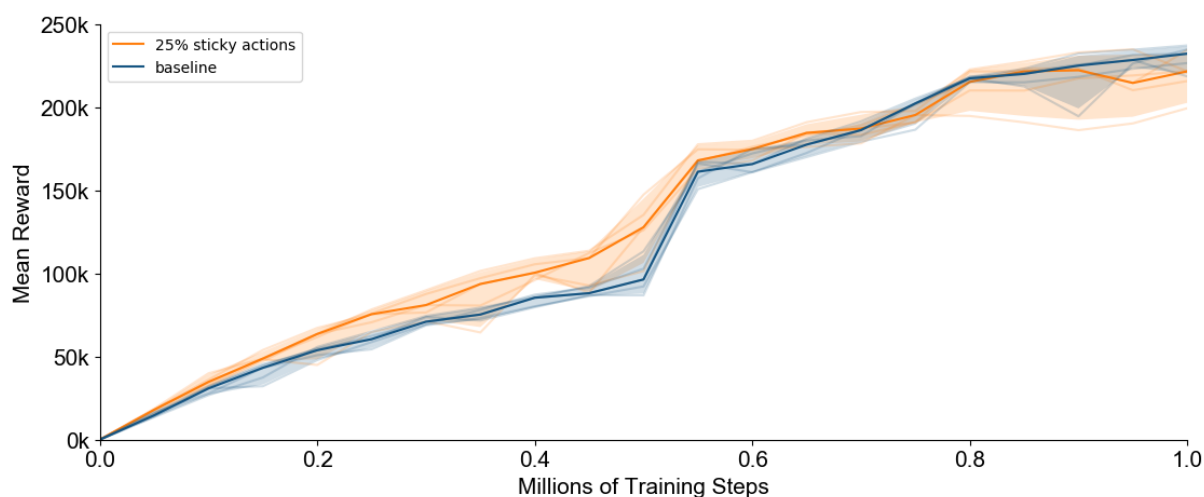


Supplementary Data Figure S3: **Details of *MuZero* evaluations (A-B) and policy improvement ablations (C-D).** (A-B) Distribution of evaluation depth in the search tree for the learned model for the evaluations in Figure 3A-B. The network was trained over 5 hypothetical steps, as indicated by the red line. Dark blue line indicates median depth from the root, dark shaded region shows 25th to 75th percentile, light shaded region shows 5th to 95th percentile. (C) Policy improvement in Ms. Pacman - a single network was trained at 50 simulations per search and is evaluated at different numbers of simulations per search, including playing according to the argmax of the raw policy network. The policy improvement effect of the search over the raw policy network is clearly visible throughout training. This consistent gap between the performance with and without search highlights the policy improvement that *MuZero* exploits, by continually updating towards the improved policy, to efficiently progress towards the optimal policy. (D) Policy improvement in Go - a single network was trained at 800 simulations per search and is evaluated at different numbers of simulations per search. In Go, the playing strength improvement from longer searches is much larger than in Ms. Pacman and persists throughout training, consistent with previous results in.⁴³ This suggests, as might intuitively be expected, that the benefit of models is greatest in precision planning domains.



Supplementary Data Figure S4: **Learning curves of *MuZero* in Atari for individual games.**

Total reward is shown on the y-axis, millions of training steps on the x-axis. Line indicates mean score across 1000 evaluation games, shaded region indicates standard deviation.



Supplementary Data Figure S5: **Performance of *MuZero* in Ms. Pacman with sticky actions.**

To ensure that *MuZero* is not taking advantage of the deterministic nature of the default Atari environment, we performed a comparison using the *sticky action*⁴ setting. In this setting, at each timestep there is a 25% chance that the selected action is ignored and that instead the previous action is repeated. Total reward is shown on the y-axis, millions of training steps on the x-axis. Dark lines indicate median scores across 5 separate training runs each, light lines indicate individual training runs, and the shaded regions indicate 25th to 75th percentile.

Game	Random	Human	SimPL ^e ₁₆	Ape-X ²⁰	R2D2 ¹⁹	<i>MuZero</i>			
						mean $\pm$	stderr	stddev	normalized
alien	227.75	7,127.80	616.90	40,805.00	229,496.90	741,812.63 $\pm$	1,634.46	51,686.25	10,747.5 %
amidar	5.77	1,719.53	74.30	8,659.00	29,321.40	28,634.39 $\pm$	25.81	816.12	1,670.5 %
assault	222.39	742.00	527.20	24,559.00	108,197.00	143,972.03 $\pm$	937.52	29,647.13	27,664.9 %
asterix	210.00	8,503.33	1,128.30	313,305.00	999,153.30	998,425.00 $\pm$	776.16	24,544.27	12,036.4 %
asteroids	719.10	47,388.67	793.60	155,495.00	357,867.70	678,558.64 $\pm$	234.46	7,414.19	1,452.4 %
atlantis	12,850.00	29,028.13	20,992.50	944,498.00	1,620,764.00	1,674,767.20 $\pm$	498.35	15,759.20	10,272.6 %
bank heist	14.20	753.13	34.20	1,716.00	24,235.90	1,278.98 $\pm$	1.10	34.93	171.2 %
battle zone	2,360.00	37,187.50	4,031.20	98,895.00	751,880.00	848,623.00 $\pm$	7,801.68	246,710.79	2,429.9 %
beam rider	363.88	16,926.53	621.60	63,305.00	188,257.40	454,993.53 $\pm$	420.25	13,289.32	2,744.9 %
berzerk	123.65	2,630.42	-	57,197.00	53,318.70	85,932.60 $\pm$	553.63	17,507.44	3,423.1 %
bowling	23.11	160.73	30.00	18.00	219.50	260.13 $\pm$	0.29	9.23	172.2 %
boxing	0.05	12.06	7.80	100.00	98.50	100.00 $\pm$	0.00	0.06	832.2 %
breakout	1.72	30.47	16.40	801.00	837.70	864.00 $\pm$	0.00	0.00	2,999.2 %
centipede	2,090.87	12,017.04	-	12,974.00	599,140.30	1,159,049.27 $\pm$	476.79	15,077.52	11,655.6 %
chopper command	811.00	7,387.80	979.40	721,851.00	986,652.00	991,039.70 $\pm$	2,937.14	92,880.58	15,056.4 %
crazy climber	10,780.50	35,829.41	62,583.60	320,426.00	366,690.70	458,315.40 $\pm$	4,718.22	149,203.12	1,786.6 %
defender	2,874.50	18,688.89	-	411,944.00	665,792.00	839,642.95 $\pm$	302.99	9,581.25	5,291.2 %
demon attack	152.07	1,971.00	208.10	133,086.00	140,002.30	143,964.26 $\pm$	8.89	281.08	7,906.4 %
double dunk	-18.55	-16.40	-	24.00	23.70	23.94 $\pm$	0.02	0.50	1,976.3 %
enduro	0.00	860.53	-	2,177.00	2,372.70	2,382.44 $\pm$	0.37	11.84	276.9 %
fishing derby	-91.71	-38.80	-90.70	44.00	85.80	91.16 $\pm$	0.09	2.77	345.6 %
freeway	0.01	29.60	16.70	34.00	32.50	33.03 $\pm$	0.01	0.35	111.6 %
frostbite	65.20	4,334.67	236.90	9,329.00	315,456.40	631,378.53 $\pm$	91.94	2,907.26	14,786.7 %
gopher	257.60	2,412.50	596.80	120,501.00	124,776.30	130,345.58 $\pm$	54.00	1,707.78	6,036.8 %
gravitar	173.00	3,351.43	173.40	1,599.00	15,680.70	6,682.70 $\pm$	7.88	249.17	204.8 %
hero	1,026.97	30,826.38	2,656.60	31,656.00	39,537.10	49,244.11 $\pm$	7.59	240.04	161.8 %
ice hockey	-11.15	0.88	-11.60	33.00	79.30	67.04 $\pm$	0.18	5.57	650.0 %
jamesbond	29.00	302.80	100.50	21,323.00	25,354.00	41,063.25 $\pm$	26.41	835.22	14,986.9 %
kangaroo	52.00	3,035.00	51.20	1,416.00	14,130.70	16,763.60 $\pm$	12.67	400.59	560.2 %
krull	1,598.05	2,665.53	2,204.80	11,741.00	218,448.10	269,358.27 $\pm$	3,229.91	102,138.88	25,083.4 %
kung fu master	258.50	22,736.25	14,862.50	97,830.00	233,413.30	204,824.00 $\pm$	32.09	1,014.83	910.1 %
montezuma revenge	0.00	4,753.33	-	2,500.00	2,061.30	0.00 $\pm$	0.00	0.00	0.0 %
ms pacman	307.30	6,951.60	1,480.00	11,255.00	42,281.70	243,401.10 $\pm$	1,306.58	41,317.61	3,658.7 %
name this game	2,292.35	8,049.00	2,420.70	25,783.00	58,182.70	157,177.85 $\pm$	497.15	15,721.20	2,690.5 %
phoenix	761.40	7,242.60	-	224,491.00	864,020.00	955,137.84 $\pm$	67.92	2,147.92	14,725.3 %
pitfall	-229.44	6,463.69	-	-1.00	0.00	0.00 $\pm$	0.00	0.00	3.4 %
pong	-20.71	14.59	12.80	21.00	21.00	21.00 $\pm$	0.00	0.00	118.2 %
private eye	24.94	69,571.27	35.00	50.00	5,322.70	15,299.98 $\pm$	0.01	0.28	22.0 %
qbert	163.88	13,455.00	1,288.80	302,391.00	408,850.00	72,276.00 $\pm$	26.72	845.10	542.6 %
riverraid	1,338.50	17,118.00	1,957.80	63,864.00	45,632.10	323,417.18 $\pm$	39.51	1,249.36	2,041.1 %
road runner	11.50	7,845.00	5,640.60	222,235.00	599,246.70	613,411.80 $\pm$	339.48	10,735.28	7,830.5 %
robotank	2.16	11.94	-	74.00	100.40	131.13 $\pm$	0.57	18.10	1,318.7 %
seaquest	68.40	42,054.71	683.30	392,952.00	999,996.70	999,976.52 $\pm$	1.27	40.03	2,381.5 %
skiing	-17,098.09	-4,336.93	-	-10,790.00	-30,021.70	-29,968.36 $\pm$	6.35	200.83	-100.9 %
solaris	1,236.30	12,326.67	-	2,893.00	3,787.20	56.62 $\pm$	16.04	507.30	-10.6 %
space invaders	148.03	1,668.67	-	54,681.00	43,223.40	74,335.30 $\pm$	207.67	6,567.19	4,878.7 %
star gunner	664.00	10,250.00	-	434,343.00	717,344.00	549,271.70 $\pm$	8,912.23	281,829.41	5,723.0 %
surround	-9.99	6.53	-	7.00	9.90	9.99 $\pm$	0.00	0.08	120.9 %
tennis	-23.84	-8.27	-	24.00	-0.10	0.00 $\pm$	0.00	0.00	153.1 %
time pilot	3,568.00	5,229.10	-	87,085.00	445,377.30	476,763.90 $\pm$	45.15	1,427.64	28,486.9 %
tutankham	11.43	167.59	-	273.00	395.30	491.48 $\pm$	0.65	20.61	307.4 %
up n down	533.40	11,693.23	3,350.30	401,884.00	589,226.90	715,545.61 $\pm$	208.75	6,601.36	6,407.0 %
venture	0.00	1,187.50	-	1,813.00	1,970.70	0.40 $\pm$	0.28	8.94	0.0 %
video pinball	0.00	17,667.90	-	565,163.00	999,383.20	981,791.88 $\pm$	3,299.54	104,340.77	5,556.9 %
wizard of wor	563.50	4,756.52	-	46,204.00	144,362.70	197,126.00 $\pm$	513.34	16,233.24	4,687.9 %
yars revenge	3,092.91	54,576.93	5,664.30	148,595.00	995,048.40	553,311.46 $\pm$	15,284.56	483,340.08	1,068.7 %
zaxxon	32.50	9,173.30	-	42,286.00	224,910.70	725,853.90 $\pm$	6,583.55	208,190.08	7,940.5 %
# best	0	5	0	5	13	37			

Supplementary Data Table S1: **Evaluation of *MuZero* in Atari for individual games with 30 random no-op starts.** Best result for each game highlighted in **bold**. Each episode is limited to a maximum of 30 minutes of game time (108k frames). *MuZero* mean $\pm$ standard error of the mean as well as standard deviation are shown for 1000 evaluation games. Mean episode returns (as reported in cited papers) are also shown for other agents. SIMPL^e was only evaluated on 36 of the 57 games, unavailable results are indicated with ‘-’. Human normalized score is

$$\text{calculated as } s_{\text{normalized}} = \frac{s_{\text{agent}} - s_{\text{random}}}{s_{\text{human}} - s_{\text{random}}}.$$

Game	Random	Human	Ape-X ²⁰	MuZero			
				mean $\pm$	stderr	stddev	normalized
alien	128.30	6,371.30	17,732.00	713,387.37 $\pm$ 4,022.57	127,204.68	11,424.9 %	
amidar	11.79	1,540.43	1,047.00	26,638.80 $\pm$ 170.02	5,376.38	1,741.9 %	
assault	166.95	628.89	24,405.00	143,900.58 $\pm$ 900.31	28,470.29	31,115.2 %	
asterix	164.50	7,536.00	283,180.00	985,801.95 $\pm$ 3,566.03	112,767.68	13,370.9 %	
asteroids	877.10	36,517.30	117,303.00	606,971.12 $\pm$ 3,088.25	97,659.03	1,700.6 %	
atlantis	13,463.00	26,575.00	918,715.00	1,653,202.50 $\pm$ 699.39	22,116.76	12,505.6 %	
bank heist	21.70	644.50	1,201.00	962.11 $\pm$ 4.06	128.25	151.0 %	
battle zone	3,560.00	33,030.00	92,275.00	791,387.00 $\pm$ 9,333.47	295,150.13	2,673.3 %	
beam rider	254.56	14,961.02	72,234.00	419,460.76 $\pm$ 813.91	25,738.04	2,850.5 %	
berzerk	196.10	2,237.50	55,599.00	87,308.60 $\pm$ 381.35	12,059.47	4,267.3 %	
bowling	35.16	146.46	30.00	194.03 $\pm$ 1.12	35.47	142.7 %	
boxing	-1.46	9.61	81.00	56.60 $\pm$ 1.24	39.22	524.5 %	
breakout	1.77	27.86	757.00	849.59 $\pm$ 2.50	78.93	3,249.6 %	
centipede	1,925.45	10,321.89	5,712.00	1,138,294.60 $\pm$ 605.71	19,154.12	13,533.9 %	
chopper command	644.00	8,930.00	576,602.00	932,370.10 $\pm$ 7,857.61	248,479.45	11,244.6 %	
crazy climber	9,337.00	32,667.00	263,954.00	412,213.90 $\pm$ 4,842.50	153,133.35	1,726.9 %	
defender	1,965.50	14,296.00	399,865.00	823,636.00 $\pm$ 358.53	11,337.84	6,663.7 %	
demon attack	208.25	3,442.85	133,002.00	143,858.05 $\pm$ 9.38	296.50	4,441.0 %	
double dunk	-15.97	-14.37	22.00	23.12 $\pm$ 0.04	1.33	2,443.1 %	
enduro	-81.84	740.17	2,042.00	2,264.20 $\pm$ 1.57	49.67	285.4 %	
fishing derby	-77.11	5.09	22.00	57.45 $\pm$ 0.41	13.12	163.7 %	
freeway	0.17	25.61	29.00	28.38 $\pm$ 0.03	1.05	110.9 %	
frostbite	90.80	4,202.80	6,512.00	613,944.04 $\pm$ 2,546.15	80,516.33	14,928.3 %	
gopher	250.00	2,311.00	121,168.00	129,218.68 $\pm$ 60.84	1,923.98	6,257.6 %	
gravitar	245.50	3,116.00	662.00	3,390.65 $\pm$ 72.96	2,307.25	109.6 %	
hero	1,580.30	25,839.40	26,345.00	44,129.55 $\pm$ 121.51	3,842.53	175.4 %	
ice hockey	-9.67	0.53	24.00	52.40 $\pm$ 0.26	8.33	608.5 %	
jamesbond	33.50	368.50	18,992.00	39,107.20 $\pm$ 64.99	2,055.05	11,663.8 %	
kangaroo	100.00	2,739.00	578.00	13,210.50 $\pm$ 107.70	3,405.72	496.8 %	
krull	1,151.90	2,109.10	8,592.00	257,706.70 $\pm$ 3,529.53	111,613.62	26,802.6 %	
kung fu master	304.00	20,786.80	72,068.00	174,623.60 $\pm$ 662.23	20,941.46	851.1 %	
montezuma revenge	25.00	4,182.00	1,079.00	57.10 $\pm$ 5.47	172.83	0.8 %	
ms pacman	197.80	15,375.05	6,135.00	230,650.24 $\pm$ 1,655.88	52,363.39	1,518.4 %	
name this game	1,747.80	6,796.00	23,830.00	152,723.62 $\pm$ 478.06	15,117.62	2,990.7 %	
phoenix	1,134.40	6,686.20	188,789.00	943,255.07 $\pm$ 620.92	19,635.33	16,969.6 %	
pitfall	-348.80	5,998.91	-273.00	-801.10 $\pm$ 37.28	1,179.00	-7.1 %	
pong	-17.95	15.46	19.00	19.20 $\pm$ 0.04	1.15	111.2 %	
private eye	662.78	64,169.07	865.00	5,067.59 $\pm$ 225.23	7,122.29	6.9 %	
qbert	159.38	12,085.00	380,152.00	39,302.10 $\pm$ 3,817.76	120,728.24	328.2 %	
riverraid	588.30	14,382.20	49,983.00	315,353.33 $\pm$ 142.19	4,496.34	2,281.9 %	
road runner	200.00	6,878.00	127,112.00	580,445.00 $\pm$ 3,492.20	110,433.05	8,688.9 %	
robotank	2.42	8.94	69.00	128.80 $\pm$ 0.57	17.92	1,938.3 %	
seaquest	215.50	40,425.80	377,180.00	997,601.01 $\pm$ 989.85	31,301.84	2,480.4 %	
skiing	-15,287.35	-3,686.58	-11,359.00	-29,400.75 $\pm$ 5.66	178.93	-121.7 %	
solaris	2,047.20	11,032.60	3,116.00	2,108.08 $\pm$ 84.66	2,677.20	0.7 %	
space invaders	182.55	1,464.90	50,699.00	57,450.41 $\pm$ 792.77	25,069.65	4,465.9 %	
star gunner	697.00	9,528.00	432,958.00	539,342.70 $\pm$ 8,963.91	283,463.86	6,099.5 %	
surround	-9.72	5.37	6.00	8.46 $\pm$ 0.03	0.87	120.5 %	
tennis	-21.43	-6.69	23.00	-2.30 $\pm$ 0.06	2.01	129.8 %	
time pilot	3,273.00	5,650.00	71,543.00	405,829.30 $\pm$ 2,408.83	76,173.96	16,935.5 %	
tutankham	12.74	138.30	128.00	351.76 $\pm$ 3.41	107.72	270.0 %	
up n down	707.20	9,896.10	347,912.00	607,807.85 $\pm$ 6,733.29	212,925.29	6,606.9 %	
venture	18.00	1,039.00	936.00	21.10 $\pm$ 1.95	61.68	0.3 %	
video pinball	0.00	15,641.09	873,989.00	970,881.10 $\pm$ 4,073.56	128,817.23	6,207.2 %	
wizard of wor	804.00	4,556.00	46,897.00	196,279.20 $\pm$ 567.00	17,930.18	5,209.9 %	
yars revenge	1,476.88	47,135.17	131,701.00	888,633.84 $\pm$ 4,790.39	151,485.44	1,943.0 %	
zaxxon	475.00	8,443.00	37,672.00	592,238.70 $\pm$ 9,941.81	314,387.59	7,426.8 %	
# best	0	6	5	46			

Supplementary Data Table S2: **Evaluation of MuZero in Atari for individual games from human start positions.** Best result for each game highlighted in **bold**. Each episode is limited to a maximum of 30 minutes of game time (108k frames). MuZero mean $\pm$ standard error of the mean as well as standard deviation are shown for 1000 evaluation games. Mean episode returns (as reported in cited papers) are also shown for other agents.