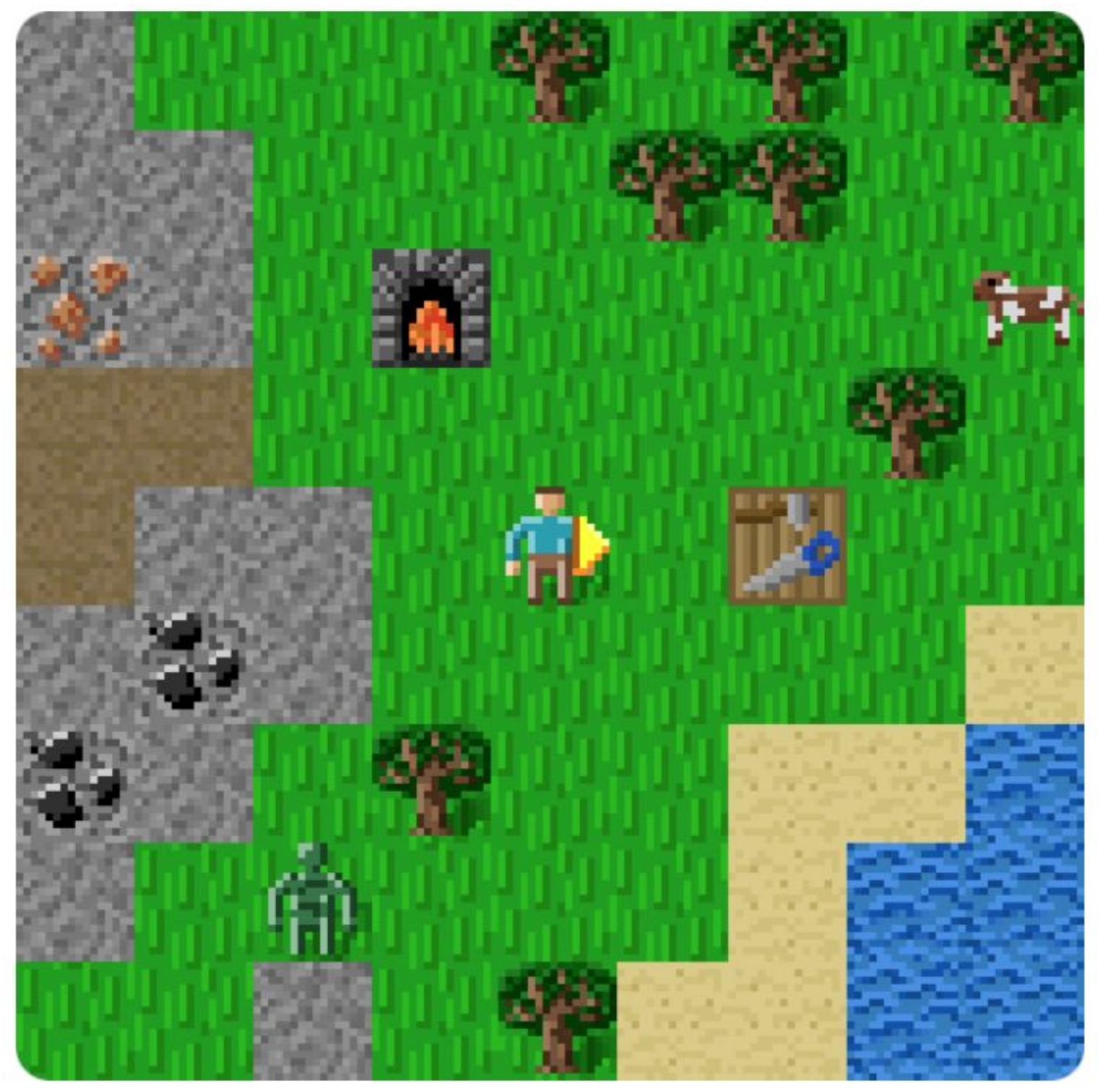
Benchmarking the Spectrum of Agent Capabilities

Danijar Hafner[1][2][3]

[1] Google Brain [2] University of Toronto [3] Vector Institute

1 Introducing Crafter

1. Crafter is an open-world survival game for RL research with and without rewards.
2. The player needs to collect resources and craft tools, all while struggling to survive.
3. The goal is to unlock 22 semantically meaningful achievements per episode.
4. Crafter evaluates a broad range of agent abilities within a single env and training run.
5. Pure Python, easy to install, use, and modify to maximize research productivity!



Agent view of a procedurally generated world in Crafter, showing terrain types, resources, and creatures.

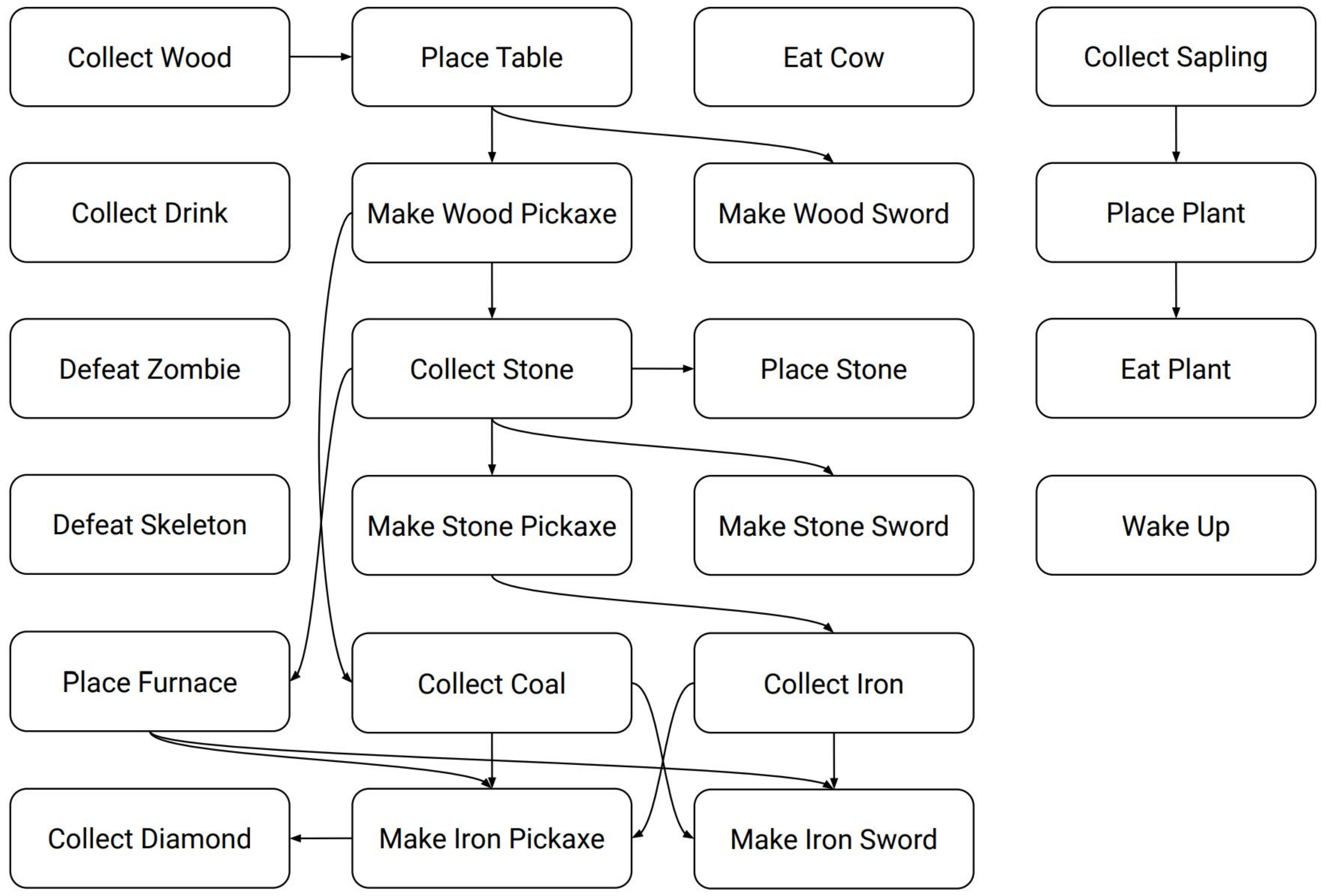
The goal of Crafter is not to replace Minecraft, but to progress towards it more quickly!

2 Research Challenges

- **Exploration** Unlocking many achievements in the technology tree requires both deep and wide exploration.
- **Generalization** Every episode uses a different procedurally generated world, so agents need to detect similar situations.
- **Reusable skills** Needs to repeatedly collect water, forage for food, and collect basic resources such as wood and stone.
- **Credit assignment** Several achievements feature long temporal dependencies, such as waiting for a plant to grow fruits to eat.
- **Memory** Agent needs to remember lakes to repeatedly find water and remember where it has already been to find rare resources.
- **Representation** Agent-centric visual observations that change with the day-night cycle require learning stable representations.
- **Survival** Having to find water, food, sleep, and defend against monsters helps prevent trivial solutions of unsupervised agents.

3 Achievements

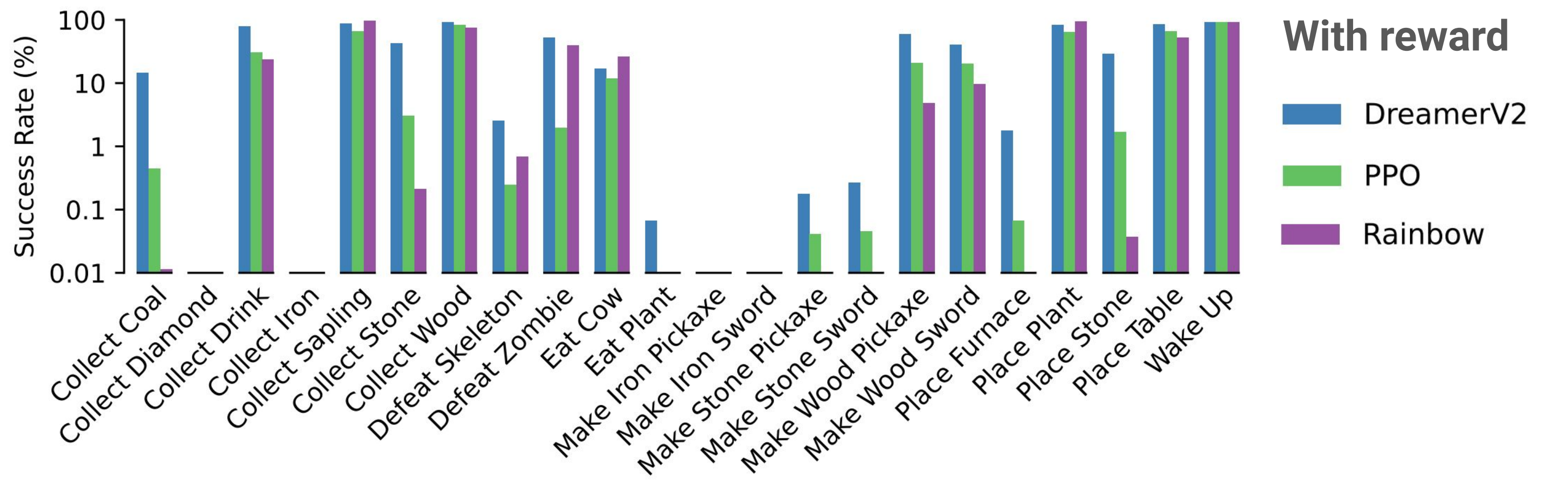
The goal of Crafter is to unlock **22 achievements per episode**, which correspond to meaningful milestones in agent behavior and measure a diverse range of abilities of agents with or without reward.

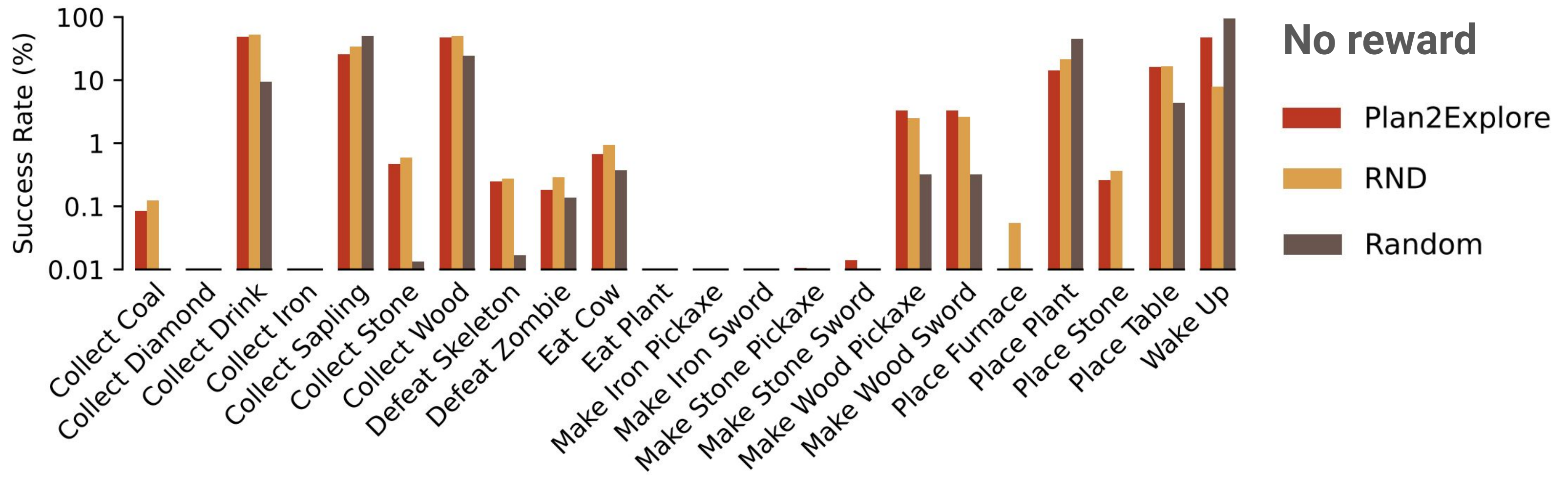


Collect Wood
Place Table
Eat Cow
Collect Sapling
Collect Drink
Make Wood Pickaxe
Make Wood Sword
Place Plant
Defeat Zombie
Collect Stone
Place Stone
Eat Plant
Defeat Skeleton
Make Stone Pickaxe
Make Stone Sword
Wake Up
Place Furnace
Collect Coal
Collect Iron
Collect Diamond
Make Iron Pickaxe
Make Iron Sword

4 Agent Ability Spectrum

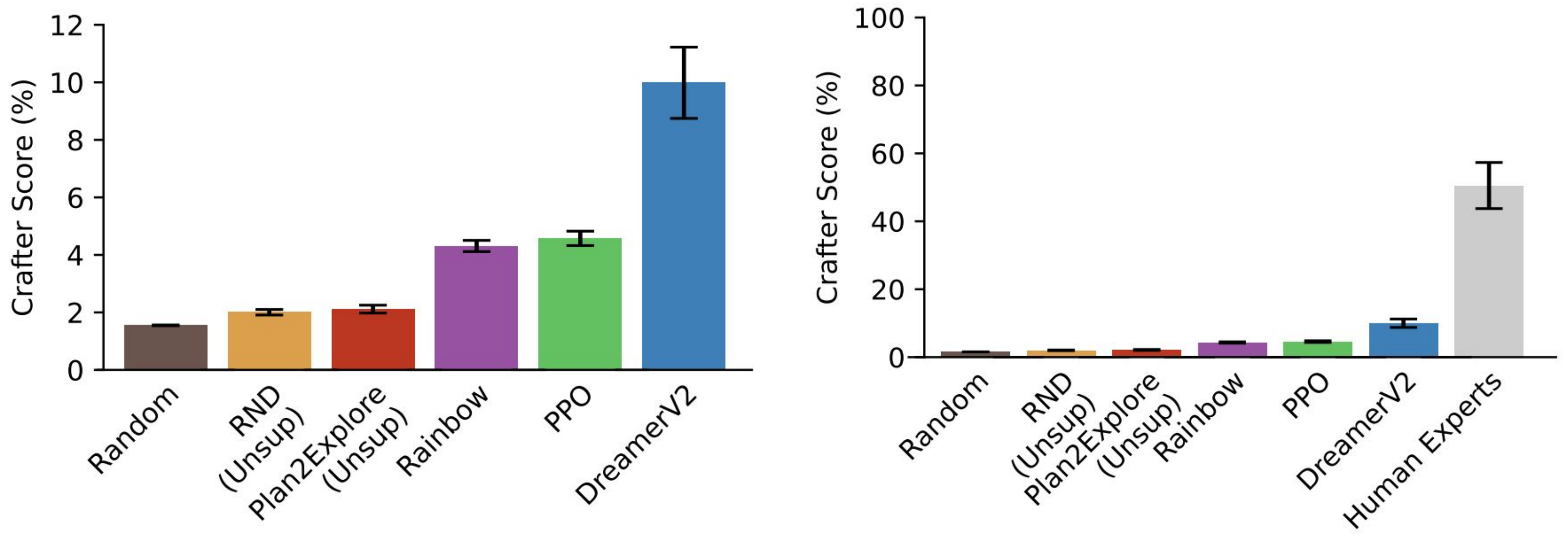
Success rates are computed per achievement across all training episodes leading up to the budget of 1M env steps. This offers insights into an agent's **strengths** and **weaknesses**.





5 Benchmark Scores

Crafter is of **appropriate difficulty**, allowing current top agents to make some learning progress while posing a substantial research challenge to reach **human performance** in the future.



(a) Crafter scores of various agents

(b) Same scores including human experts

The benchmark score is the **geometric mean** over success rates:

This is useful when tasks are of different difficulty or reward scale, without having to know or assume the difficulties a priori.

Geometric mean (mean in log-space) weighs difficult tasks stronger, rewarding agents for demonstrating broad abilities over many tasks.

Collecting diamonds 1% of the episodes instead of 0% is a big improvement. Collecting wood 95% instead of 90% is not.

6 Available Resources

Play Crafter Yourself!

Supports Linux, Mac, Windows

```
$ python3 -m pip install crafter   # Install Crafter
$ python3 -m pip install pygame    # Needed for human interface
$ python3 -m crafter.run_gui       # Start the game
```

Useful resources are available on the project website:

danijar.com/crafter

- Human expert video
- Emergent agent behaviors
- Baseline implementations (Docker)
- Baseline scores (JSON)
- Plotting scripts
- Human dataset (NPZ)

Achievement	Human Experts
Collect Coal	86.0%
Collect Diamond	12.0%
Collect Drink	92.0%
Collect Iron	53.0%
Collect Sapling	67.0%
Collect Stone	100.0%
Collect Wood	100.0%
Defeat Skeleton	31.0%
Defeat Zombie	84.0%
Eat Cow	89.0%
Eat Plant	8.0%
Make Iron Pickaxe	26.0%
Make Iron Sword	22.0%
Make Stone Pickaxe	78.0%
Make Stone Sword	78.0%
Make Wood Pickaxe	100.0%
Make Wood Sword	45.0%
Place Furnace	32.0%
Place Plant	24.0%
Place Stone	90.0%
Place Table	100.0%
Wake Up	73.0%
Score	50.5%

@danijarh

Website with videos, data, and code: danijar.com/crafter