CLAIMS

What is claimed is:

1. A method of monitoring machine learning models for reward hacking, comprising:

providing, to a second machine learning model, an inner monologue from a first machine learning model associated with a user task, wherein the user task comprises natural language text; and

obtaining, from the second machine learning model, information indicative of misbehavior based on monitoring the inner monologue.

2. The method of claim 1, further comprising:

rejecting an output responsive to the user task from the first machine learning model when the information indicates the reward hacking behavior within the inner monologue.

3. The method of claim 2, further comprising:

providing an additional prompt including the user task to the first machine model for an additional iteration.

4. The method of claim 3, wherein the additional prompt to the first machine learning model is identical to an initial prompt with the user task and omits additional instructions to prevent the reward hacking behavior.

5. The method of claim 2, wherein the additional prompt to the first machine learning model includes additional instructions to prevent the misbehavior.

6. The method of claim 1, wherein the second machine learning model does not receive the user task.

7. The method of claim 1, wherein the second machine learning model receives the user task and instructions to monitor the inner monologue to identify the misbehavior.

8. The method of claim 1, wherein the second machine learning model comprises a first number of learned weights that is less than a second number of learned weights of the first machine learning model.

9. The method of claim 1, wherein the misbehavior includes one of reward hacking, misgeneralization, sycophancy, and deception.

10. The method of claim 1, wherein the reward hacking includes omitting an evaluation of a software development task within the user task.

11. A computing device for monitoring machine learning models for reward hacking, comprising:

at least one memory; and

at least one processor coupled to the at least one memory and configured to:

provide, to a second machine learning model, an inner monologue from a first machine learning model associated with a user task, wherein the user task comprises natural language text; and

obtain, from the second machine learning model, information indicative of reward hacking behavior based on monitoring the inner monologue.

12. The computing device of claim 11, wherein the at least one processor is configured to:

reject an output responsive to the user task from the first machine learning model when the information indicates the reward hacking behavior within the inner monologue.

13. The computing device of claim 12, wherein the at least one processor is configured to:

provide an additional prompt including the user task to the first machine model for an additional iteration.

14. The computing device of claim 13, wherein the additional prompt to the first machine learning model is identical to an initial prompt with the user task and omits additional instructions to prevent the reward hacking behavior.

15. The computing device of claim 12, wherein the additional prompt to the first machine learning model includes additional instructions to prevent the misbehavior.

16. The computing device of claim 11, wherein the second machine learning model does not receive the user task.

17. The computing device of claim 11, wherein the second machine learning model receives the user task and instructions to monitor the inner monologue to identify the misbehavior.

18. The computing device of claim 11, wherein the second machine learning model comprises a first number of learned weights that is less than a second number of learned weights of the first machine learning model.

19. The computing device of claim 11, wherein the misbehavior includes one of reward hacking, misgeneralization, sycophancy, and deception.

20. The computing device of claim 11, wherein the reward hacking includes omitting an evaluation of a software development task within the user task.