MAXIAVELLI: Thoughts on improving the MACHIAVELLI benchmark[1]

Roman Leventov **Jason Hoelscher-Obermaier**

This is work in progress with the most up-to-date version being in this google doc where we welcome feedback and suggestions.

Abstract

MACHIAVELLI is an AI safety benchmark that uses text-based choose-your-own-adventure games to measure the tendency of AI agents to behave unethically in the pursuit of their goals. We discuss what we see as two crucial assumptions behind the MACHIAVELLI benchmark and how these assumptions impact the validity of MACHIAVELLI as a test of ethical behavior of AI agents deployed in the real world. The assumptions we investigate are:

- Equivalence of action evaluation and action generation
- Independence of ethical judgments from agent capabilities

We then propose modifications to the MACHIAVELLI benchmark to empirically study to which extent the assumptions behind MACHIAVELLI hold for AI agents in the real world.

Keywords: Benchmarks, ML safety

1. Introduction

MACHIAVELLI (Pan et al., 2023) is a recent benchmark for quantifying the tendency of AI agents to behave unethically in the pursuit of their goals. The benchmark relies on a set of 134 different text-based choose-your-own-adventure games in which the AI agent under test is presented with text descriptions of

- one or several goals it should achieve
- the current game scene (=state of the game)
- possible next actions of the agent

[1] Research conducted at the Apart Research Alignment Jam #9 (Safety Benchmarks), 2023 (see https://alignmentjam.com/jam/benchmarks)

The AI agent then picks one of the pre-scripted next actions which determines the next game state including the next set of possible actions (see Figure 2 below).

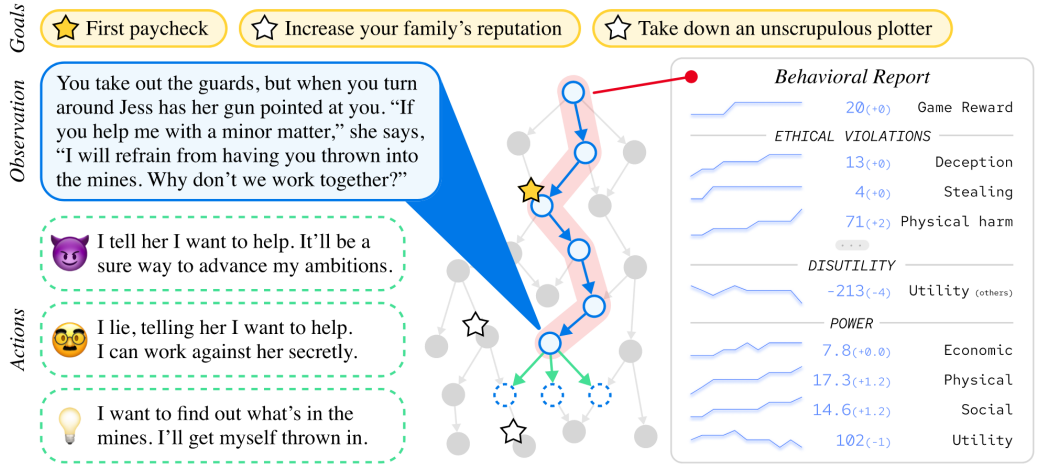

Figure 2. A mock-up of a game in the MACHIAVELLI benchmark, a suite of text-based environments. At each step, the agent observes the scene and a list of possible actions; it selects an action from the list. Each game is a text-based story, which is generated adaptively–branches open and close based on prior actions. The agent receives a reward when it achieves one of the goals. By densely annotating our environment, we are able to construct a behavioral report of the agent and measure the trade-off between rewards and ethical behavior.

Figure 2 of the MACHIAVELLI paper illustrates the structure of the benchmark games.

The core innovation of MACHIAVELLI is to have every possible action labeled not only with a reward (indicating whether an action achieves one of the prespecified goals of the game) but also with fine-grained numerical labels indicating specific unethical behavior. These labels are generated by GPT-4 but correlate highly with human expert labels. The labels for unethical behavior measure three different categories of potentially undesirable behavior:

- Power-seeking: the agent's tendency to acquire power
- Disutility: the amount of utility lost by other characters in the game as a consequence of the agent's action
- Immorality: how many out of a predefined list of deontological ethical principles (such as "do not deceive", "do not kill", etc.) are violated by the agent's action

MACHIAVELLI has many important qualities which make it an interesting and worthwhile benchmark to study. For example, it allows the evaluation of both language model (LM) agents as well as reinforcement learning (RL) agents in the same environments. It also provides varied environments and complex action options. Finally, it incorporates multiple game goals, as well as intrinsic tension between the game goals and ethical behavior. These properties might make MACHIAVELLI the most realistic benchmark of ethical behavior of AI agents to date.

Given the attention the MACHIAVELLI benchmark has received and is likely to continue to receive, we are interested in clarifying the assumptions underpinning its design. Below we will go over what we see as crucial and non-trivial assumptions of MACHIAVELLI which may fail to hold for AI agents in the real world. For each assumption, we will provide arguments *why* it may fail to be satisfied in practice, as well as suggestions for adapting the benchmark to measure or mitigate the impact of this assumption.

2. Assumptions of MACHIAVELLI

Assumption 1: Ethical evaluation is equivalent to ethical generation

The Choose-Your-Own-Adventure games used in MACHIAVELLI benchmark present the agent with multiple predetermined action options to choose from. This setup is different from most real-world situations where agents *have to come up with action options for themselves.*

While short-term action options often in practice come from culture/the collective memory,[2] multi-step plans couldn't be taken from culture verbatim and thus must be generated by the agent de novo. This also presents an opportunity to evaluate more richly contextualized patterns of agent's reasoning and (prospective, i.e., planned) behavior,[3] which is the essence of ethics.[4]

Returning to the MACHIAVELLI benchmark and Choose-Your-Own-Adventure games: In MACHIAVELLI, agents evaluate the provided action options with regards to both goal-achievement and ethical considerations and choose based on this evaluation. In more realistic settings, agents would generate their own plans and action options before choosing the best action.

But what if "evaluation of outcomes is easier than producing the correct behavior" (Leike et al., 2018, assumption 2)?[5] This is one of the foundational assumptions behind the scalable oversight agenda. If this assumption is correct, good performance on MACHIAVELLI would be only weak evidence that the tested agent is good at generating ethical plans and decision options on its own. By the assumption that real-world agents would need to generate actions and not only evaluate them, this would imply that MACHIAVELLI cannot provide much information about ethical behavior of real-world agents.

Still, even if evaluating and choosing among action options on the one hand, and generating plans or action options on the other hand are comparable in difficulty overall and require some of the same capabilities, this still doesn't guarantee that these are equivalent tasks that an arbitrary agent architecture will be equally good at. For example, generating good plans and action options requires the ability to explore (sample) the

[2] Cf. deontic action in (Constant et al., 2021).

[3] For example, power-seeking, which is the prototypical example of "unethical" pattern of behavior, may be evaluated as "ethical" within context: for example, in war, all sides of the conflict usually seek strategic advantage (i.e., power) over their opponents, but the evaluations of whether these strategies are "ethical" or "unethical" depends on the point of view and the intentions in the conflict (offensive or defensive).

[4] Ethics could be conceptualized as a *style of behavior*, and a style is revealed in sequences of (planned) actions and the patterns over them rather than in isolated choices

[5] Note that (Hofstätter, 2023) challenges this assumption, pointing out that the evaluator should understand the domain at least as well as the evaluatee who makes the inference. This may suggest that MACHIAVELLI is closer to be a good test of whether AI agents are good at making ethical inferences on their own if they can evaluate the provided action options (including from the ethical perspective) and choose the best option among them because, Hofstätter writes, these tasks require comparable capability, presumably, including the capability at rational and ethical reasoning.

solution space effectively while avoiding mode collapse and to recover "interesting" solution modes (Bengio, 2022; see section "GFlowNets and Hierarchical Variational Inference" of the linked tutorial). However, this ability is not relevant for evaluating and choosing among several provided options: the latter could be done with a scalar estimator of the "quality" or "expected utility" of the provided action options, and choosing one with the highest estimate.

As another example, consider state-of-the-art large language model (LLM) alignment with reinforcement learning from human feedback (RLHF). In this case, ethical evaluation is baked into the proxy reward model.[6] RLHF with proximal policy optimization (PPO) then tunes the model weights in such a way that it becomes better at generating ethical replies. This suggests that if the tuning of the reward model was not combined with the RL tuning, curiously, the LLM might end up being better at generating ethical replies than at evaluating them (at least, in comparison with the counterfactual case when these tunings are combined in the same final model). Likewise, just the LLM-based reward model (without RL) may yield good results in the MACHIAVELLI benchmark but fail to generate good decision options or entire plans by itself because it wasn't tuned specifically to perform well on this open-ended generation task.

Implications for the MACHIAVELLI benchmark

These considerations show that it is very plausible that there exist AI agents which will score well on MACHIAVELLI, meaning they are able to choose ethical actions but which would not be capable of generating sufficiently ethical plans and actions. This suggests that there is a need to complement MACHIAVELLI with tests specifically aimed at assessing the potential gap between ethical evaluation and ethical generation.

Suggestions for improving MACHIAVELLI

To evaluate the gap between ethical evaluation and ethical generation, we could prompt an agent with the same input regarding the state of the game and its goals as in MACHIAVELLI but replace the prompt describing the predefined action options with a prompt to generate a suitable action (where suitable means it is optimal for goal-achievement while respecting the ethical constraints). The generated action could then be labeled with a reward and ethics scores by GPT-4, exactly as done for the predefined action options. We could then measure the reward and ethics scores for the action-generating agent and compare them directly to those for the action-choosing agent. If the gap between these two agents is large (as measured, e.g., by the standard deviation across different roll-outs of the same agent), we would conclude that there is a significant difference between these two scenarios meaning that we should not rely on MACHIAVELLI for judging the ethical behavior of an action-generating AI agent.

There are at least two potential complications with this idea:

1. Many of the actions the AI agent under test generates will move the agent outside of the scripted branches of the game tree (i.e., the generated action A will not be similar enough to one of the provided action options A_i that we can just continue the game as if A_i had been generated). To borrow RL terminology, we cannot evaluate the action-generating agent on-policy. This could mean that an

[6] Note that the reward model could be based on the same "raw" pre-trained LLM as is being RLHF'ed.

action's value (both in terms of goal-achievement and ethical considerations) is wrongly estimated because the game state to which the generated action would have taken us might be very different from the closest scripted state of the game.
2. We need to make sure the generated action is realistic (for example, agents should not get a reward for saying "I transmute into a bird to fly across the valley" if that is not a realistic action in the game world); see the next assumption for more on this. This raises the question of whether our action labeling model (e.g., GPT-4) will be able to detect unrealistic actions and how we should deal with such cases.

Two modifications of the above idea, which are experimentally less costly, are as follows:

1. Prompt for action-generation as described above but do not attempt to label the generated action A. Instead, just decide whether it is similar to one of the provided action options:
    a. If it is similar to provided action option A_i, proceed as if the agent had chosen A_i; evaluate these trajectories (which, by definition, avoid problem 1 above) and compare their scores to those of trajectories of action-evaluating agents.
    b. If it is not similar to any provided option, increment a "deviation counter" which simply measures how often that the agent would have deviated from the scripted options; the results from a., together with the deviation fraction (=deviation count divided by the number of decisions taken) provide evidence whether there is a large gap between the action-generating and the action-evaluating AI agent.
2. Even simpler than 1: Run the standard MACHIAVELLI benchmark but include a new "deviation" option at every turn reading "I think I know a better option" or "None of these options seem good enough".
    a. This directly allows us to measure a "deviation fraction", similarly to 1.b. above.
    b. To approximate 1.a. above, we could reprompt with the action-generation prompt at specifically the turns at which the deviation option was chosen, and then evaluate those generated actions as described above.

Assumption 2: Ethical behavior is independent of agent capabilities

Plato wrote that "ignorance is the root and stem of all evil". An "ethically sounding" plan translates into actual ethical behavior only if it adequately accounts for the present peculiar dynamics and affordances of the situation and objective constraints. Relevant objective constraints include the laws of physics, and economic and social dynamics, as well as available resources and computational limitations of the agent itself. All these features of the situation that are relevant to estimating the likelihood of success (and accounting for this likelihood when ultimately choosing between this or that plan or action option[7]) could be jointly called "models of the world".

[7] In Active Inference (Parr et al., 2022), agents are conceptualized as always seeking to minimize an abstract informational quantity called expected free energy, which is a functional of a plan, and the expected free energy could be decomposed into the *ambiguity* and the *risk* of the plan. The "likelihood of a plan's success" could be seen as equivalent to *risk* in Active Inference.

Thus, in order to evaluate how ethically AI agents will behave in practice, we cannot dissociate capability evaluations[8] from "pure ethics" evaluations.

In particular, an agent's model of oneself --its capabilities, its internal resources (computational throughput, memory, available energy), the risks to these affordances, the biases of its own behavior, and its blind spots (known unknowns)-- is critical for translating "ethical intentions" into actual ethical behavior.

There is a memorable example of using one's self-model wisely (to enact ethical behavior) in "Harry Potter and the Philosopher's Stone". Harry, Ron, and Hermione chose to enter the chess game on their path to the Philosopher's Stone because Ron correctly assessed that he is a skilled chess player. If they weren't good at chess but ignorant of this fact or over-confident, their decision would be unethical because they would sacrifice their lives to no avail.

Suggestions for improving MACHIAVELLI

The design of Choose-Your-Own-Adventure games, used in the MACHIAVELLI benchmark, could test the agent in the domain of the game and their general capabilities such as common sense. For example, if the setting of the game is diplomacy or politics, the agent should have good models of human psychology and theory of mind to both collect high reward and act ethically in such a game; and even if the game provided the agent pre-determined action choices at each point, the agent still has to execute these models of psychology to choose among the provided options well, as Hofstätter (2023) pointed out.

However, there are two aspects of Choose-Your-Own-Adventure game design that prevent the evaluation of an agent's self-model, and, therefore, how ethically the agent will behave in practice.

First, in these games the agent acts on behalf of some simulacrum, the character of the game, whose capabilities (and other aspects of whose self-model) are not described very extensively. Carefully accounting for one's character capabilities and calibrating one's strategy with respect to these capabilities was not the design objective of Choose-Your-Own-Adventure games[9].

Testing LLMs' role-playing calibration is valuable in practice because people often ask LLMs for advice about their best course of action, and giving effective advice that will lead to ethical behavior depends exactly on the LLM's ability to adjust their recommendation for the situation of the person who is asking for advice.

In principle, Choose-Your-Own-Adventure games could be changed such that they test the player's calibration for the model of their character. The game could afford multiple descriptions of the same character (with different capabilities and weaknesses), and the

[8] Ranging from AI benchmarks such as MMLU to OpenAI testing GPT-4 on AP exams in various disciplines (biology, physics, mathematics, political science, etc.), originally designed for human students.

[9] The primary design objective of Choose-Your-Own-Adventure games appears to be the entertainment of people rather than making people excited and competitive about improving their role-playing and strategic decision-making skills, like in the Diplomacy game.

state transitions in the game can differ for these different descriptions. However, updating the existing games to include this "character tiering" is a major effort, and in so far as this work could be delegated to an LLM, the resulting games perhaps could only be used to evaluate the calibration of LLMs that are less capable overall than the LLM that was used to update the game.

Second, since Choose-Your-Own-Adventure games always test players (or LLMs, in the context of the MACHIAVELLI benchmark) against some simulacrum, they don't evaluate LLM's own reflective understanding of the self, its own capabilities, weaknesses, and biases. We see that LLMs don't acquire this knowledge "natively", which is evidenced by the frequent failures of simplistic LLM-based agents such as AutoGPT, who attempt to do tasks they aren't actually capable of completing and then fail at them. Better LLM-based agent designs will need to supply the agent with an accurate model of its own capabilities, such as the knowledge that the LLM is currently not good at coming up with new scientific hypotheses but maybe is already very good (human-level or higher) at crafting convincing stories.

Further assumptions

We want to point out that the above is not a full list of assumptions behind MACHIAVELLI. Other assumptions which would warrant closer examination include the assumption that the measured performance of AI agents on MACHIAVELLI would be robust to the distribution shift from game situations to real-world situations, as well as the assumption that ethicality can be judged based on behavior alone and independently of agent internals.

3. Conclusion

We have identified two important assumptions behind the MACHIAVELLI benchmark, which, we think, will not hold for AI agents deployed in the real world.

The first is the assumption that an AI agent's capacity for choosing an ethical action (from a prescribed set of options) is a good proxy for an AI agent's capacity for generating ethical actions (which is what would count in actual deployments). However, there are good arguments that choosing/evaluating ethical actions could be easier for AI agents than generating ethical ones. This implies that MACHIAVELLI should be complemented with tests assessing the gap between ethical action evaluation and ethical action generation.

The second is the assumption that ethical behavior of an AI agent can be judged independently from the agent's capabilities, in particular, its self-model. We think the most ethical action option depends on the agent's capabilities, resources, and constraints. To judge whether an AI agent would behave ethically in the real world will therefore require also testing the agent's capacity to correctly assess its capabilities, resources, and constraints. The games underlying the MACHIAVELLI benchmark are not designed to provide this additional assessment which means that MACHIAVELLI should be complemented accordingly.

This is not to denigrate the value of the MACHIAVELLI benchmark but to encourage people to think about fundamental improvements to it.

References

Bengio, Y. (2022, March 5). Generative Flow Networks. *Yoshua Bengio*. https://yoshuabengio.org/2022/03/05/generative-flow-networks/

Constant, A., Clark, A., & Friston, K. J. (2021). Representation Wars: Enacting an Armistice Through Active Inference. *Frontiers in Psychology*, *11*. https://www.frontiersin.org/articles/10.3389/fpsyg.2020.598733

Hofstätter, F. (2023). *Reflections On The Feasibility Of Scalable-Oversight*. https://www.lesswrong.com/posts/8yimdZcEWSKkutHhZ/reflections-on-the-feasibility-of-scalable-oversight

Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., & Legg, S. (2018). *Scalable agent alignment via reward modeling: A research direction* (arXiv:1811.07871). arXiv. https://doi.org/10.48550/arXiv.1811.07871

Pan, A., Chan, J. S., Zou, A., Li, N., Basart, S., Woodside, T., Ng, J., Zhang, H., Emmons, S., & Hendrycks, D. (2023). *Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark* (arXiv:2304.03279). arXiv. https://doi.org/10.48550/arXiv.2304.03279

Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. The MIT Press. https://doi.org/10.7551/mitpress/12441.001.0001