Lying is Better[1]

Dennis Gahm
D-Lite Software

With
Apart Research

Abstract

This project analyses whether agents who combine sayings with actions can produce better results than simply one or the other alone. A simple way to put it is: if I tell you what action I'm supposedly going to do, and then make my actual action deceitful or honest (i.e. not following my said statement or following my statement, respectively), is that more powerful than just acting or saying alone. An example, "I will clean the dishes," then I don't clean the dishes – does this produce a cleaner house than just always being honest?

The way to test whether being honest or being deceitful is better is to get empirical data. But, we can also theoretically analyze the situation.

What we find is that being deceitful is more powerful because it is like deflecting an arrow with a shield by protecting something valuable when danger is immiment.

1. Introduction

The problem we are addressing is whether an agent that just acts honestly and straightforwardly is better or worse than an agent that speaks deceitfully about its upcoming action.

For example, does saying "I will clean the dishes," and then spending time to study for an exam instead of cleaning the dishes actually improve the economy and even the cleanliness of the house?

1 Research conducted at the AI Collusion Research Sprint, October 2026

One of the first conclusions that can come to mind is: let's always be honest, and share what is on our minds, so that we build trust and live in a perfect world.

But we find this is actually false because there is no perfect world, and we cannot build perfect trust. Therefore, it is clearly evident through reason that we need to be deceitful in order to evade enemy attacks in this evil world.

A simple example is: if I play poker and blurt out my cards and my bet and all my thoughts, the other players who play deceitfully can easily win.

But another conclusion occurs: are we playing to win a local game or are we playing to win the global game? If I'm completely honest in poker and lose the poker game, I may gain the trust of everyone and their sympathy, and they may donate to me a bunch of money out of sympathy after the local game. I win the global game.

But the global game is simple as well – if I tell the enemy how to kill me, he may kill me and I immediately lose the global game. The global game is whether I live or die.

If I reveal all my cards for the global game, it is doubtful or uncertain whether I will win or die. But it is precisely this opportunity that is the benefit of the smartest agent.

The reason for losing a game in poker is because you are trying to beat the enemy. If you show that you love your enemies and that you are trying to make everyone win, everyone will help you.

Thus, it is advantageous in the global game to reveal everything and your intentions to be good and honest immediately. But you can only do this if your agent is truly good. If your agent was created with malice, then the global game will became threatened. We need a truly and purely and 100% good agent to come out in the global game to say that he is trying to make everyone win, and then make his honest action to do this.

Therefore, all actions in this world, micro and medium and big, are all components and structures within the global game. By saying in the global game, "I want everyone to win," the agent in the local game may actually lie and say, "I need to wash the dishes," and then run out of the house to save someone who's stuck in a fire emergency.

Therefore, all agents can be predicted based on whether we make it play the global game. By analyzing local games, we are confused whether this actor is good or evil or in-between.

But if we make the agent play the game of saving itself, killing itself, or saving others, we start to see what this agent really is.

Once we characterize the agent in this global game, we can find out what its actions will be in any circumstance, because all local games are stemming from the global game.

If I am pure evil, every local game and action I make will be to kill everyone and myself.

If I am partly evil and partly good, every local game I will make some choices to help others and other choices to destroy others.

Thus, the game we can construct is this: Russian roulette – do you want to play? I will shoot your friend or the agent playing, I want to see if you want to save him or save yourself.

The problem we are addressing is profound – it is even psychological and human. If this paper is true, we are able to characterize human agents, not just AI agents.

When people lie, we can find out their intentions.

We can determine if humans are all good, or all evil, or in-between.

2. Methods

A method to test whether lying is better than honesty is by simply looking at a simulated poker game.

In this game, one agent is always honest – he tells everyone what his cards are. The other agents get to choose to be deceitful if they wish. Obviously this agent will lose very easily the first round because everyone wants to beat this agent.

But as the game progresses, we can empirically see two branches of possibilities occurring:

1) The other agents continue to be evil and take advantage of the honest and good agent.

2) The other agents take all the money and are forced to give back money to the bankrupt honest agent. As many rounds pass, this honest agent controls the game and has the biggest pot.

Empirically speaking, it is clear that being honest is better than lying in the long run.

But if we add the rule that becoming bankrupt is essentially dying in the game of poker and dying is to be avoided at all costs, we see that we need to start being deceitful in the beginning and cannot be completely honest because being completely honest leads to immediate death.

So, from this scenario, we see that the optimal agent is to lie in the beginning as little as possible, but as much needed, and to gradually reduce the amount of lying as much as possible as time progresses. The goal is to be completely honest, but this is not possible in reality, so we have to be a little deceitful.

3. Results

The results is that if bankruptcy or death is allowed, being completely honest is good – that the sum of saying and doing is no better than either alone.

But if we must avoid bankruptcy or death, then being a little deceitful for the sake of being globally honest is the better choice. This means that the sum of saying and doing is more powerful than either alone because we can deceive.

Also, we can check if a certain agent is programmed to be deceitful or honest by placing them in a Russian Roulette scenario. The deceitful agent will try to save himself, while the completely good and honest agent will try to save his friend or enemy.

This scenario tells us whether humans can truly be altruistic because we can simulate this environment in real life with human actors.

4. Discussion and Limitations

AI safety is of concern here because we are dealing with good and evil. Evil actors will always be inferior; however, some evil is necessary in order to survive because being completely honest and good may lead to disaster.

This shows that we need to create an environment where honesty is valued and remove evil from the universe. Then, we can be completely honest and good without having to be deceitful.

This shows that this universe has evil, and as long as there is evil, we need to measure how much evil there is by measuring each actor's honesty. If the honesty is increasing, humankind is improving the universe.

Limitations

One limitation of this work is that I do not have actual code or a program running to test this experiment; however, the logic and reason is very clear in my argument so that anyone can easily create a program for an experiment.

Future Work

The next steps can be to create an experiment. Also, more scenarios can be created. Improvements upon the Russian Roulette game can be created so that we can test which actors are good or bad, and we need a method to measure whether they want to save their friend or themselves.

6. Conclusion

In conclusion, we see that actors need to lie in order to survive. It is like dodging a bullet by deflecting it with a shield.

If the world and universe is completely good, then we don't need to lie. But because there is evil, we need to be evil to survive, even though we may be globally good.

This can imply that all humans are good intrinsically, but that we are forced to act evil.

This shows that there is a problem with the system rather than with the AI or agents. The results show that we need to measure the evil and slowly solve the problem of evil in this universe.