Which Games are Unaffected by Absolute Commitments?

Extended Abstract

Daji Landis
Bocconi University
Milan, Italy
daji@ucla.edu

Nikolaj I. Schwartzbach
Bocconi University
Milan, Italy
nikolaj@ignatieff.io

ABSTRACT

We identify a subtle security issue that impacts mechanism design in scenarios in which agents can absolutely commit to strategies, by which the strategy of an agent may depend on the commitments made by the other agents. This changes fundamental game-theoretic assumptions by inducing a meta-game of choosing which strategies to commit to. We say that a game that is unaffected by such commitments is *Stackelberg resilient* and show that computing it is intractable in general, though it can be computed efficiently for two-player games of perfect information. We show the intuitive, but technically non-trivial result, that if a game is resilient when some number of players can make commitments, it is also resilient when these commitments are available to fewer players. We demonstrate the non-triviality of Stackelberg resilience by analyzing two escrow mechanisms from the literature. These mechanisms have the same intended functionality, but we show that only one is Stackelberg resilient. Our model is particularly relevant in Web3 scenarios, where these commitments can be realized by the automated and irrevocable nature of smart contracts, and highlights an important issue in ensuring the secure design of Web3. In particular, our work suggests that smart contracts already deployed on major blockchains may be susceptible to these attacks.

KEYWORDS

Stackelberg resilience; mechanism design; smart contracts

ACM Reference Format:
Daji Landis and Nikolaj I. Schwartzbach. 2024. Which Games are Unaffected by Absolute Commitments?: Extended Abstract. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 ABSOLUTE COMMITMENTS

We introduce the concept of an *absolute commitment*, where agents have a more 'absolute' ability to commit to strategies than is usually the case in games. In particular, we grant agents the capacity to make irrevocable commitments that can condition on the content of other agents' strategies. For commitments to be irrevocable, the strategy, once chosen, cannot be altered by the agent. We will generally consider these strategies to be specified using self-executing programs. The ability of commitments to condition on the contents of other commitments makes sense in a context, such as with smart contracts in Web3, in which all commitments are both self-executing and knowable to other agents.

The *letters of last resort* are secret letters, written by an incoming British prime minister to the commanders of submarines. They stipulate what the commanders should do in the case that a nuclear strike destroys the British Government. In particular, they could call for nuclear retaliation, even the possibility of which should act as deterrence. If we assume compliance by the commanders, we have an example of a self-executing strategy. It is not subgame perfect to enact meaningless retaliation, so a letter ordering such destruction illustrates a change in equilibrium as compared to a game without irrevocable commitments. This point was a tenet of mutually assured destruction — a Cold War deterrence strategy — that was predicated on the idea that, if one country were annihilated, their submarines would annihilate the aggressor, despite it being too late for such action to save the homeland. While this tells part of the story we explore in this paper, there is an aspect missing. To fully encapsulate our narrative, we first require that the letter be known to the potential aggressor and that the aggressor could condition their strategy based on the strategy it contains. In this version, the incoming British prime minister might be told,

> "*Our spies tell us that our enemy has ordered their submarines that, if you write a letter ordering retaliation in the case of strike, then their submarine commanders are to carry out a strike! We know that enemy spies will know the content of your letter with complete certainty.*"

The incoming prime minister now finds themselves in a bind: a letter stipulating retaliation will precipitate an attack and certain retaliation, while a letter forbidding an attack will leave Britain vulnerable. In this story we can already make two observations: there is a decided first mover advantage, Britain here must respond to commitments already made by an enemy; and what was before a protracted game of many moves will now be decided solely in the commitment phase, after which the agents can only sit back and watch the system unfold, knowing precisely what will happen.

This idea of a leader-follower dynamic is captured by a Stackelberg competition model. In particular, games in which agents can commit to strategies are known as *Stackelberg games* [12] and their equilibria are known to be hard to compute [1, 7]. In this model, the leader can do backwards induction to predict what a follower might do in response to their plan. In our scenario, rather than the leader simply being able to deduce what the followers will do in response to their actions, the leader tells the followers precisely what will be done in response to any possible follower strategy. A regular Stackelberg model returns us to the simple example wherein a leader says that they will retaliate should destruction befall their homeland. Having made the inference that no follower would then

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

dare attack, the leader can rest assured they will not need their threat. A leader in the case of our absolute commitments would have to follow a different strategy. If the incoming British PM were the leader and were not responding to an enemy, they may declare, '*If the enemies of Britain stipulate no retaliation, so will we. Otherwise we will preemptively attack.*' Such a commitment would clearly lead to the followers complying with the threat and all the submarine crews could go home. These more complex equilibria are known as reverse Stackelberg equilibria [2, 3] and are also studied in a variety of contexts, such as in control theory [4, 5, 11].

In this work, we consider extensive-form games and use subgame perfect equilibria (SPE) as solution concepts. We will use the model of absolute commitments proposed by Hall-Andersen and Schwartzbach [6], which strictly generalizes (reverse) Stackelberg equilibria. In this model, granting an agent the capacity to commit absolutely corresponds to allowing this agent to make a 'cut' in the game tree, which must respect information sets. This induces an 'expanded game' of exponential size, containing a root node belonging to the agent and a subgame corresponding to each possible cut for the agent. For multiple commitments, the commitments are expanded in a bottom-up manner, which gives a natural means for these commitments to condition on each other. Hall-Andersen and Schwartzbach show that reasoning about the subgame equilibria of these games generalizes (reverse) Stackelberg games and is hard in the general case.

Given the prevalence of known, algorithmically stipulated games, we find this both a practically important and an interesting line of questioning. One clear potential environment for these absolute commitments is in smart contracts [6]. Smart contracts are decentralized programs that run on a virtual machine implemented by a blockchain, such that, once deployed, their execution is no longer under the control of their creator [13]. They are generally used to store and allocate funds, providing clear economic incentives for these attacks. Importantly, in most blockchains, the contents of smart contracts are public and can, in principle, reason about each other[1]. This setting was studied by Landis and Schwartzbach in the context of blockchain transaction fee mechanisms [8]. In their work, a group of agents have transactions that they want to include on a block, the contents of which is controlled by a miner. The agents then pay the miner to have their transactions included on the block. However, [8] demonstrates that the leading agent may commit absolutely such that their transaction is included at zero cost, and force the other agents to enter into a lottery for the remaining space on the block. In that instance, the outcome benefited all the agents, except for the miner, and the threats were largely just adhering to the regular SPE, whereas the commitments allowed the agents to spontaneously collude. By contrast, in this work, we show instances in which the introduction of additional contracts is favorable for only one or a few of the agents.

1.1 Our Results

As mentioned, games may in some contexts be susceptible to attacks in which agents commit to self-executing strategies that change the nature and equilibrium of the game. We call these attacks *Stackelberg attacks*, since reasoning about these attacks captures Stackelberg games as a special case. We observe that some games are resilient to these attacks, in the sense that the set of subgame perfect equilibria is unchanged by adding sequential commitments for the players (in any order). We say that such a game is *Stackelberg resilient*. We now state the main results of this work; proofs of all these statements can be found in the full version of this paper [9].

We first investigate the computational complexity of determining if a game is Stackelberg resilient. We show, using techniques in [6], that Stackelberg resilience is hard to compute in general, but can be computed efficiently in some simple cases.

Theorem 1 (Computational Complexity). *Determining Stackelberg resilience is PSPACE-hard in general, although it can be determined efficiently for two-agent games of perfect information.*

Technically, what we show is that Stackelberg 1-resilience is NP-hard for games of imperfect information, using the same reduction as in [6]. However, this result can be extended to show PSPACE-hardness when the number of agents is unbounded (regardless of whether the games have perfect or imperfect information).

Next, we analyze two escrow mechanisms from the literature [**? ?**]. These two mechanisms have the same intended functionality: namely, holding a payment in escrow while a trade is being finalized. Interestingly, we find that one of these mechanisms is indeed Stackelberg resilient, while the other one is not.

Theorem 2 (Non-Triviality). *Stackelberg resilience is non-trivial: there are two escrow mechanisms, only one of which is resilient.*

Essentially, in one of the games the seller may create a self-executing strategy that forces the buyer to dispute delivery of an item they actually received (causing the buyer to lose their deposit). This demonstrates that Stackelberg resilience is a non-trivial property and begs the question of which mechanisms can be implemented in a Stackelberg resilient manner. We leave it as interesting future work to develop techniques to design mechanisms that are Stackelberg resilient.

Finally, a natural question is whether games retain Stackelberg resilience when one agent's ability to make absolute commitments is taken away, i.e. whether k-resilience implies $(k-1)$-resilience. We call this property *downward closure* and show that it holds in general for Stackelberg resilience.

Theorem 3 (Downward Closure). *If a game is Stackelberg resilient when k agents can make self-executing strategies, it is also Stackelberg resilient when ℓ agents have this capacity for any $\ell \leq k$.*

We show this by showing the contrapositive statement: a game that is not $(k-1)$-resilient cannot be k-resilient. The proof also implies a monotonicity property: once an agent has a viable attack, that attack cannot be undermined by adding an additional self-executing strategy when that commitment is the final one.

ACKNOWLEDGMENTS

NIS was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 101019547). Part of the work was done while NIS was a Ph.D. student at Aarhus University.

[1]Note that while Rice's theorem [10] states that any non-trivial property of Turing machines is undecidable, in threatening an agent into deploying contract X, a smart contract need not check for semantic equivalence, only whether or not the agent deploys exactly contract X.

REFERENCES

[1] Vincent Conitzer and Tuomas Sandholm. 2006. Computing the Optimal Strategy to Commit To. In *Proceedings of the 7th ACM Conference on Electronic Commerce* (Ann Arbor, Michigan, USA) *(EC '06)*. Association for Computing Machinery, New York, NY, USA, 82–90. https://doi.org/10.1145/1134707.1134717

[2] Noortje Groot, Bart De Schutter, and Hans Hellendoorn. 2012. Reverse stackelberg games, part i: Basic framework. In *2012 IEEE International Conference on Control Applications*. 421–426.

[3] Noortje Groot, Bart De Schutter, and Hans Hellendoorn. 2012. Reverse Stackelberg games, part II: Results and open issues. In *2012 IEEE International Conference on Control Applications*. IEEE, 427–432.

[4] Noortje Groot, Bart De Schutter, and Hans Hellendoorn. 2014. Toward system-optimal routing in traffic networks: A reverse stackelberg game approach. *IEEE Transactions on Intelligent Transportation Systems* 16, 1 (2014), 29–40.

[5] Noortje Groot, Georges Zaccour, and Bart De Schutter. 2017. Hierarchical game theory for system-optimal control: Applications of reverse Stackelberg games in regulating marketing channels and traffic routing. *IEEE Control Systems Magazine* 37, 2 (2017), 129–152.

[6] Mathias Hall-Andersen and Nikolaj I. Schwartzbach. 2021. Game Theory on the Blockchain: A Model for Games with Smart Contracts. In *Algorithmic Game Theory*, Ioannis Caragiannis and Kristoffer Arnsfelt Hansen (Eds.). Springer International Publishing, Cham, 156–170.

[7] Dmytro Korzhyk, Vincent Conitzer, and Ronald Parr. 2010. Complexity of computing optimal stackelberg strategies in security resource allocation games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 24. 805–810.

[8] Daji Landis and Nikolaj Schwartzbach. 2023. *Stackelberg Attacks on Auctions and Blockchain Transaction Fee Mechanisms*. https://doi.org/10.3233/FAIA230501

[9] Daji Landis and Nikolaj I. Schwartzbach. 2023. Which Games are Unaffected by Absolute Commitments? arXiv:2305.04373 [cs.GT]

[10] Henry Gordon Rice. 1953. Classes of Recursively Enumerable Sets and Their Decision Problems. *Trans. Amer. Math. Soc.* 74, 2 (1953), 358–366. http://www.jstor.org/stable/1990888

[11] Mohammad Amin Tajeddini, Hamed Kebriaei, and Luigi Glielmo. 2020. Decentralized hierarchical planning of PEVs based on mean-field reverse stackelberg game. *IEEE Transactions on Automation Science and Engineering* 17, 4 (2020), 2014–2024.

[12] Heinrich von Stackelberg. 1934. *Marktform und Gleichgewicht*. Verlag von Julius Springer.

[13] Gavin Wood. 2014. Ethereum: A secure decentralised generalised transaction ledger. *Ethereum project yellow paper* 151 (2014), 1–32.