

КОМИСИЈИ ЗА СТУДИЈЕ II СТЕПЕНА ЕЛЕКТРОТЕХНИЧКОГ ФАКУЛТЕТА У БЕОГРАДУ

Комисија за студије II степена, Електротехничког факултета у Београду, на својој седници одржаној 11.09.2025 године именовала нас је у Комисију за преглед и оцену мастер рада дипл. инж. Немања Савески под насловом „Супарнички напади у учењу подстицањем вођени имитацијама“. Након прегледа материјала Комисија подноси следећи

ИЗВЕШТАЈ

1. Биографски подаци кандидата

Немања Савески је рођен 07.09.2000. године у Шапцу. Завршио је основну школу „Николај Велимировић“ у Шапцу као вуковац. Уписао је Шабачку гимназију коју је такође завршио као вуковац. Током школовања такмичио се на свим нивоима из математике, физике и програмирања. Електротехнички факултет уписао је 2019. године. Дипломирао је на одсеку за Сигнале и системе 2023. године са просечном оценом 8,80. Дипломски рад “Стеганографија слике заснована на конволуционим аутоенкодерским неуралним мрежама” одбранио је у јулу 2023. године са оценом 10. Дипломске академске – мастер студије на Електротехничком факултету у Београду, на Модулу за сигнале и системе уписао је у октобру 2023. године. Положио је све испите са просечном оценом 9,40. У марту 2024. уписао је и мастер студије Наука о подацима (ен. Data Science) на Факултету за Информатику Техничког Универзитета у Бечу.

2. Извештај о студијском истраживачком раду

Кандидат Немања Савески је као припрему за израду мастер рада урадио истраживање релевантне литературе која се односи на области учења подстицањем, учења имитација и безбедности у машинском учењу. У оквиру прегледа, посвећена је пажња супарничким нападима на алгоритме учења подстицањем где нападач нема приступ политици циљних агената, при чему се у већини радова углавном разматрају напади измена улазних стања. За разлику од тога, у овом раду су од већег значаја били напади засновани на измени акција, док су претходно наведени, напади измена улазних стања, коришћени само као упоредна основа. Током истраживања уочен је феномен који је у раду описан као парадокс имитација и пророка (где су пророци нападачи који имају приступ политици циљних агената), што је послужило као смерница за даље разматрање и усмеравање рада.

3. Опис мастер рада

Мастер рад обухвата 36 нумерисаних страна са укупно 17 слика, 6 табела и 38 референци. Рад садржи резиме, увод, 5 поглавља и закључак (укупно 7 поглавља) и списак коришћене литературе.

Прво поглавље представља увод у коме је дефинисан предмет истраживања и мотивисан значај безбедности у учењу подстицањем.

У другом поглављу је дат преглед основа учења подстицањем, учења имитација као и безбедности у машинском учењу.

У трећем поглављу је детаљно представљен метод рада – обучавање циљних агената, прикупљање демонстрација (понашања) као јединог извора информација о циљном агенту, тренирање имитација методом DQfD (ен. Deep Q-learning from Demonstrations), ограничење и испитивање окружења на основу имитација, као и предложени оквир за нападаче засноване на изменама акција и стања.

Четврто поглавље описује окружење у коме су извршени експерименти као и хипер-параметре свих циљних агената и нападача коришћених у методама напада.

У оквиру петог поглавља су представљени резултати који обухватају криве тренинга, табеле и евалуацију циљних агената, имитација, нападача заснованих на изменама акција, као и резултати напада заснованих на изменама стања.

Шесто поглавље, дискусија, је подељено на две целине. У првој целини је спроведена анализа ентропије политике и истакнуто је раздвајање два фактора унутар јединственог механизма напада – тајминга интервенције и потенцијала штете интервенције – што је послужило као објашњење феномена у коме пророци показују слабији ефекат у односу на нападаче који експлоатишу имитације. У другој целини су приказане могућности и усмерења за будући рад.

У седмом поглављу је дат закључак који сумира главне резултате са посебним освртом на уочени

парадокс имитација и пророка као и импликације за будуће истраживање и развој одбрамбених техника.

4. Анализа са кључним резултатима

Мастер рад дипл. инж. Немање Савеског се бави проблематиком безбедности у учењу подстицањем, са посебним освртом на супарничке нападе у условима ограничених информација где се имитације користе као инструмент за експлоатацију апроксимираних информација о циљним агентима.

Основни резултати рада су: 1) Формулисање оквира за супарничке нападе засноване на изменама акција који експлоатишу имитације као и метода за нападе измена стања где имитације служе као извор градијената; 2) откривање и формулисање парадокса имитација и пророка, који показује да нападачи који експлоатишу имитације могу да надмаше пророке; 3) увођење квантилног ограничења над укупном наградом нападача у епизодним окружењима која немају природну горњу и доњу границу укупне награде; 4) анализа механизма који доводе до уочених резултата, са посебним нагласком на улогу ентропије политике.

5. Закључак и предлог

Кандидат Немања Савески је у свом мастер раду успешно развио и применио нови оквир за анализу супарничких напада у учењу подстицањем. Рад је показао да раздвајање фактора интервенције унутар јединственог механизма пружа објашњење уочених феномена. Приказани резултати имају значајне импликације за разумевање безбедности модела учења подстицањем.

Кандидат је исказао самосталност и систематичност у своме поступку као и иновативне елементе у решавању проблематике овог рада.

На основу изложеног, Комисија предлаже Комисији за студије II степена Електротехничког факултета у Београду да рад дипл. инж. Немање Савеског прихвати као мастер рад и кандидату одобри јавну усмену одбрану.

Београд, 03.10.2025. године

Чланови комисије:

др Предраг Тадић Ванредни професор
сагласан, 03.10.2025.

Марија Новичић Асистент
сагласан, 03.10.2025.