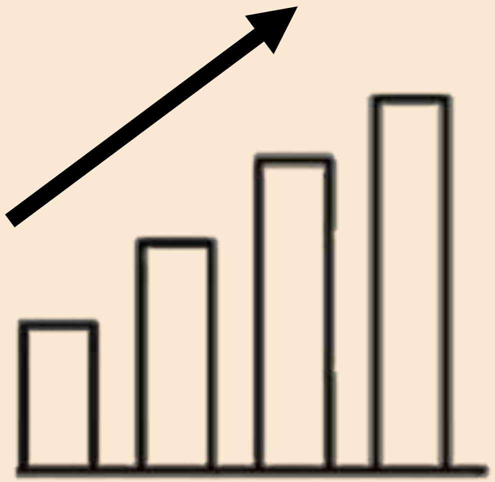
ავტორი: ელისაბედ ვარშანიე
ხელმძღვანელი: ათინა თოიძე

ანდაცვა

სად გამოოიყენება ხელოვნური ინტელექტი

ჯანდაცვა

ფინანსები

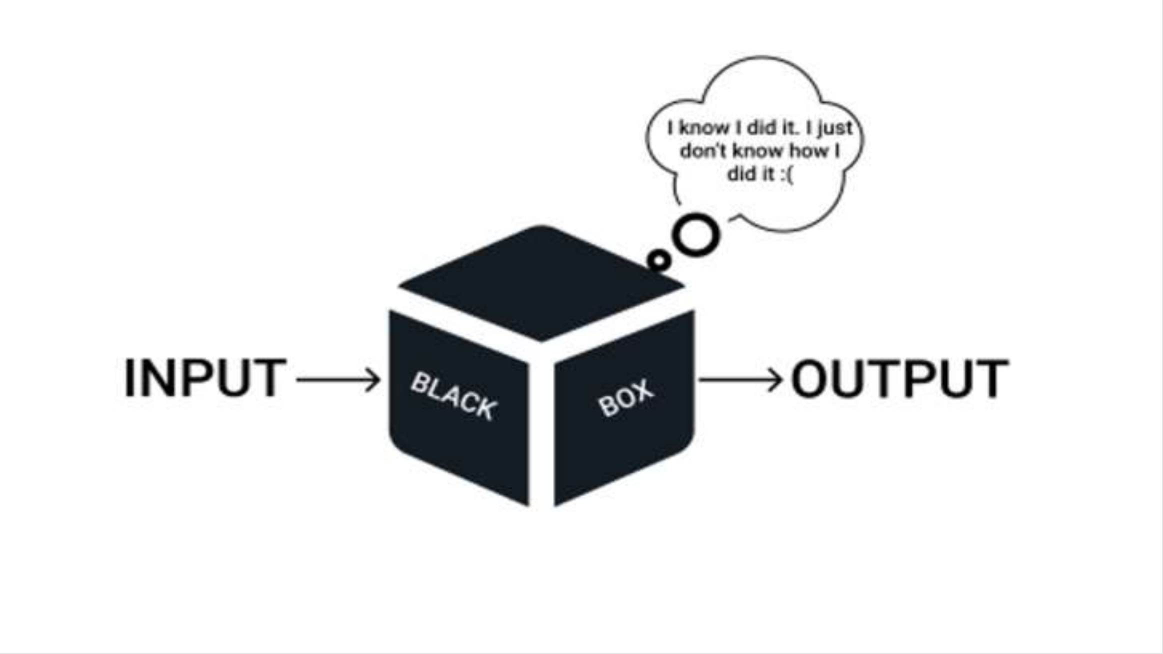


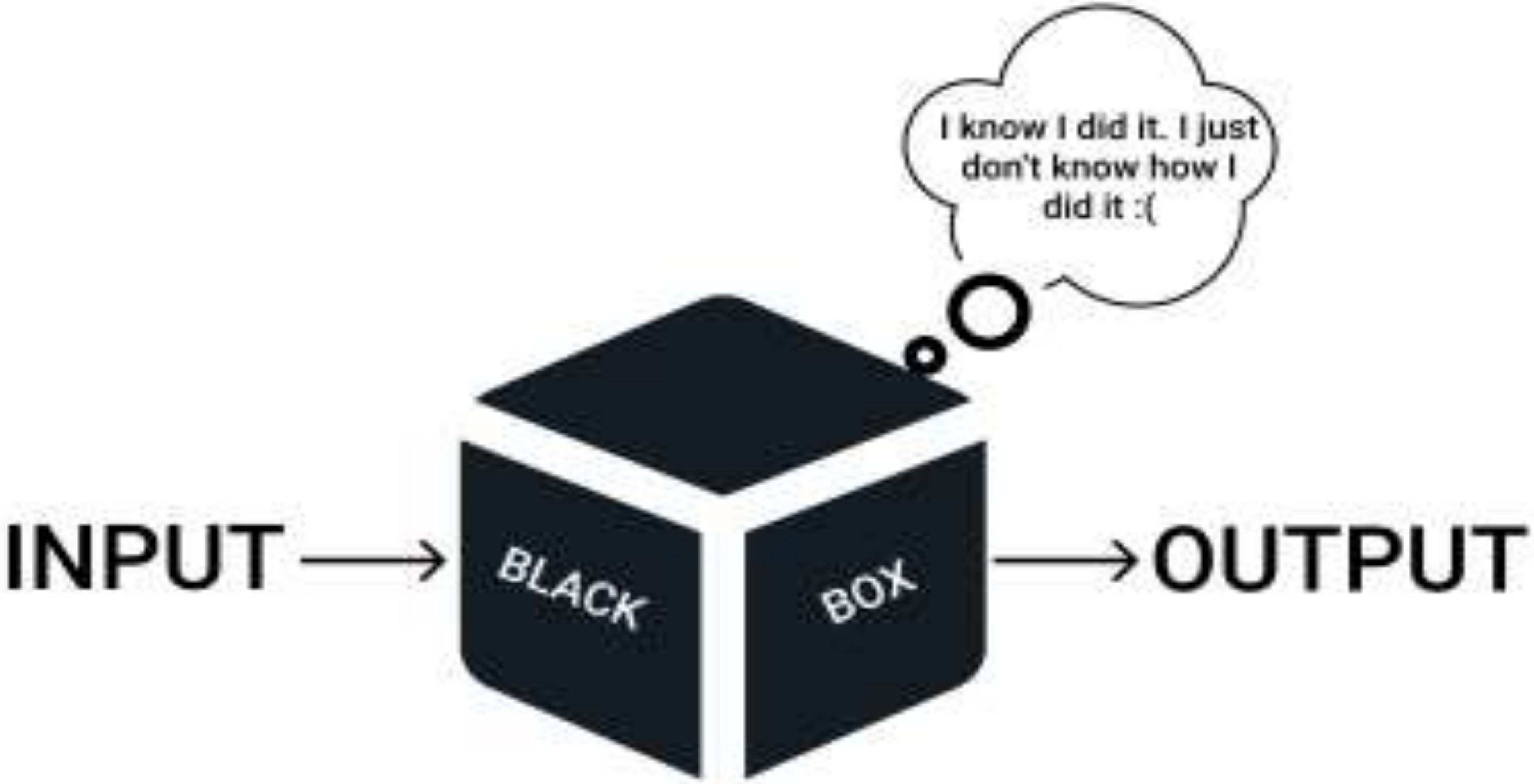
სამართალი

ყოველდღიურ ცხოვრებაში

რა არის შავი ყუთის პრობლემა?

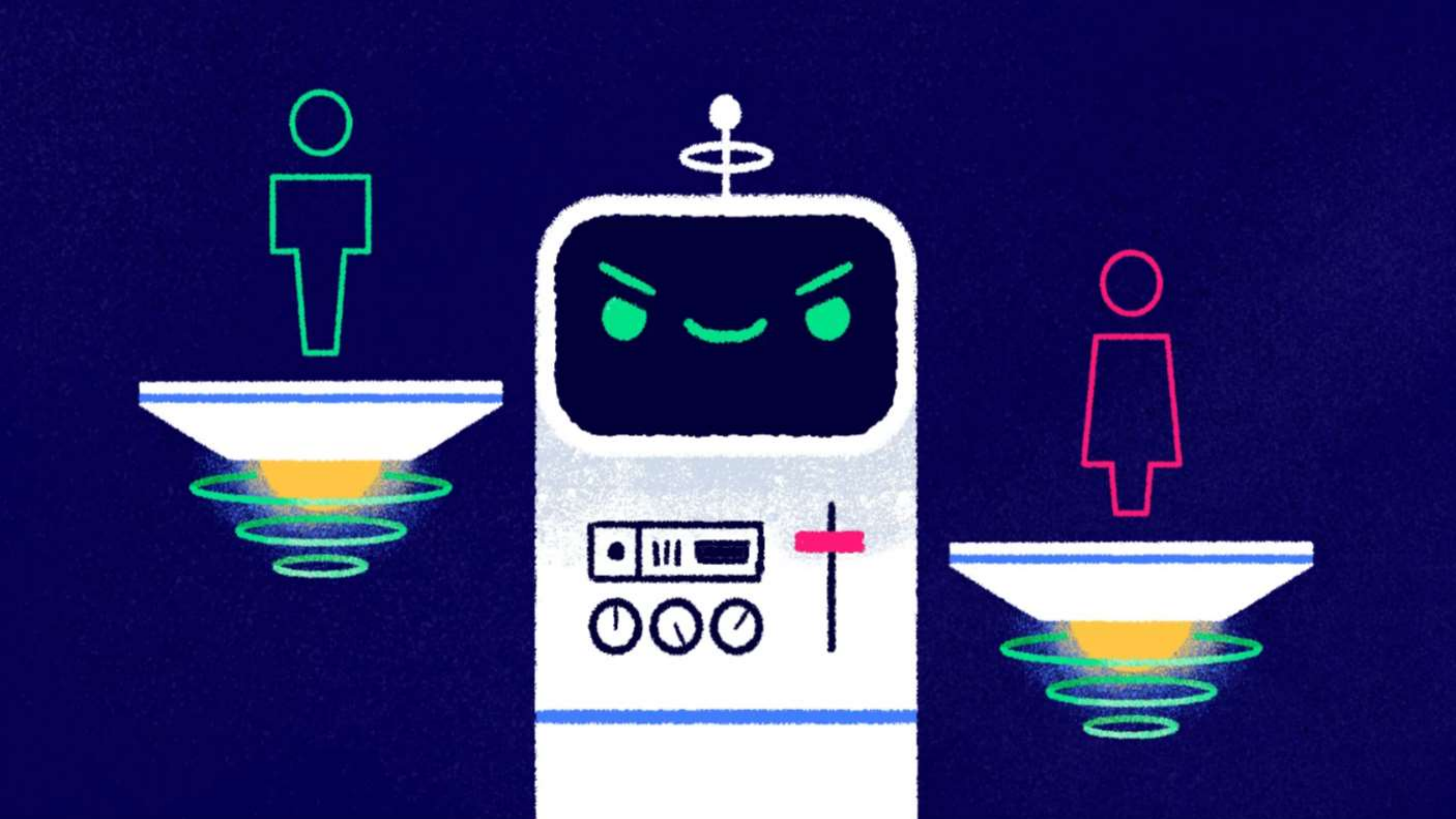
შავი ყუთის პრობლემა ხელოვნურ ინტელექტში ნიშნავს იმას, რომ ზოგიერთ სისტემაში არ ვიცით, როგორ იღებს AI გადაწყვეტილებებს. ისინი ამუშავებენ დიდ რაოდენობის მონაცემს და იღებენ საბოლოო შედეგს ისე, რომ შიდა პროცესი ადამიანისთვის გაუგებარი რჩება.
ამიტომ მას შავ ყუთს (Black Box) ვეძახით — თითქოს შიგნით რაღაც ხდება, მაგრამ ზუსტად რა, არ ვიცით. ეს პრობლემა სერიოზულია ისეთ სფეროებში, როგორიცაა მედიცინა, სამართალი, ფინანსები და თვითმართვადი მანქანები, სადაც მნიშვნელოვანია ვიცოდეთ, რატომ მიიღო სისტემამ კონკრეტული გადაწყვეტილება.
ამ პრობლემის გადასაჭრელად მეცნიერები მუშაობენ იმაზე, რომ შეიქმნას ამხსნელი AI — მოდელები, რომლებიც არა მხოლოდ სწორ პასუხს გვთავაზობენ, არამედ ხსნიან, როგორ მივიდნენ მასამდე.




I know I did it. I just don't know how I did it :(
INPUT
BLACK
BOX
OUTPUT

ალგორითმული დისკრიმინაცია

- **ალგორითმული დისკრიმინაცია** არის ფენომენი, როდესაც კომპიუტერული ალგორითმები, განსაკუთრებით მონაცემებზე დაყრდნობით მომუშავე ხელოვნური ინტელექტის სისტემები, იწვევს არასამართლიან ან მიკერძოებულ გადაწყვეტილებებს ინდივიდების ან ჯგუფების მიმართ ისეთი თვისებების საფუძველზე, როგორიცაა სქესი, რასა, ასაკი, ენა, ეთნიკური კუთვნილება, სექსუალური ორიენტაცია ან სხვა სოციალური ნიშნები.
- ეს დისკრიმინაცია, როგორც წესი, წარმოიშობა მაშინ, როდესაც ალგორითმი სწავლობს წარსულ მონაცემებზე, რომლებიც თავად შეიცავს ადამიანური მიკერძოებების, სტრუქტურული უთანასწორობის ან სისტემური დისკრიმინაციის შედეგებს. შედეგად, სისტემა იმეორებს ან აძლიერებს ამ მიკერძოებას, მიუხედავად იმისა, რომ მისი მიზანი თითქოსდა ობიექტური და ნეიტრალური გადაწყვეტილებების მიღებაა.
- ალგორითმული დისკრიმინაცია შეიძლება გამოიხატოს როგორც პირდაპირი (მაგალითად, როცა მოდელი აშკარად გამოარჩევს ადამიანებს გარკვეული ნიშნით), ისე არაპირდაპირი გზით (როდესაც გადაწყვეტილება ეფუძნება ცვლადებს, რომლებიც დაუფარავად უკავშირდება დაცულ ნიშნებს). მიუხედავად იმისა, რომ ასეთი ალგორითმები ტექნიკურად ეფექტურია, ისინი ხშირად არღვევენ ეთიკურ და სამართლებრივ პრინციპებს, როგორიცაა თანასწორობა, სამართლიანობა და მიუკერძოებლობა.



მადლობთ მოსმენისთვის<3

- თუ გსურთ უფრო მეტი გაიგოთ ამ საკითხზე შეგიძლიათ იხილოთ სრული ნაშრომი: