Fostering Trust on ML Inferences

Dalmo Cirne – November 2023

Trust me (Why?)

Trust me (Why?)

Because the ML Model Said So

Trust me (Why?)

~~Because the ML Model Said So~~

Not a Good Explanation

Explainability

Classifications	Confidence Score

Explainability

Classifications			Confidence Score
Classification 1	Classification 2	Classification 3	0.10

Explainability

Classifications			Confidence Score
Classification 1	Classification 2	Classification 3	0.10
—	Classification 2	Classification 3	0.15

Explainability

Classifications			Confidence Score
Classification 1	Classification 2	Classification 3	0.10
—	Classification 2	Classification 3	0.15
Classification 1	—	Classification 3	0.92

Explainability

Classifications			Confidence Score
Classification 1	Classification 2	Classification 3	0.10
—	Classification 2	Classification 3	0.15
Classification 1	—	Classification 3	0.92
Classification 1	Classification 2	—	0.39

Explainability

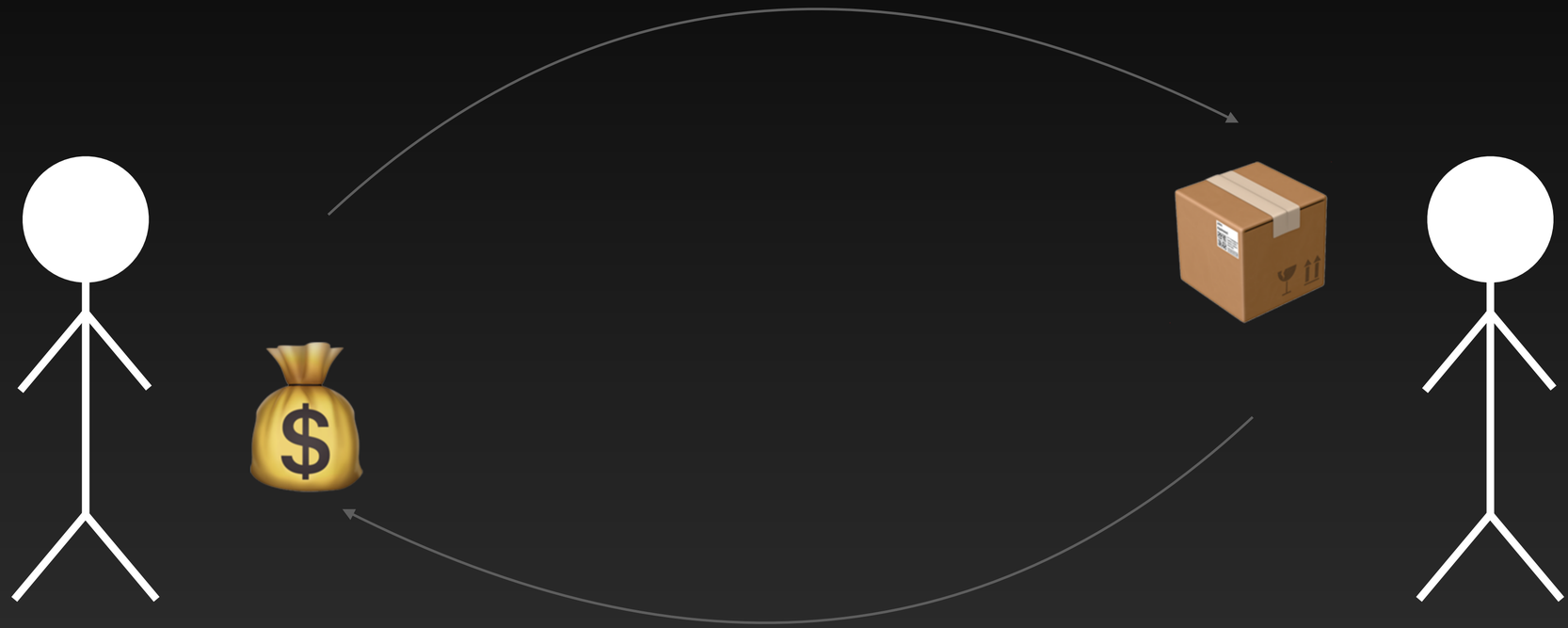
Classifications			Confidence Score
Classification 1	Classification 2	Classification 3	0.10
—	Classification 2	Classification 3	0.15
Classification 1	—	Classification 3	0.92
Classification 1	Classification 2	—	0.39

Explainability

Classifications			Confidence Score
Classification 1	Classification 2	Classification 3	0.10
—	Classification 2	Classification 3	0.15
Classification 1	—	Classification 3	0.92
Classification 1	Classification 2	—	0.39

Exchanging Value

Exchanging Value



Game Theory

Chess

Game Theory

Chess

Prisoner's Dilemma

	Talk	Don't Talk
Talk		
Don't Talk		

Game Theory

Chess

Prisoner's Dilemma

	Talk	Don't Talk
Talk		
Don't Talk		-2, -2

Game Theory

Chess

Prisoner's Dilemma

		Talk	Don't Talk
	Talk		
	Don't Talk	-12, 0	-2, -2

Game Theory

Chess

Prisoner's Dilemma

	Talk	Don't Talk
Talk		0, -12
Don't Talk	-12, 0	-2, -2

Game Theory

Chess

Prisoner's Dilemma

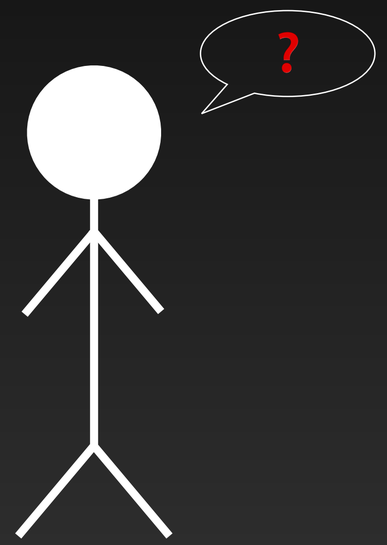
		Talk	Don't Talk
	Talk	-7, -7	0, -12
	Don't Talk	-12, 0	-2, -2

Trust

Trustworthy

Not Trustworthy

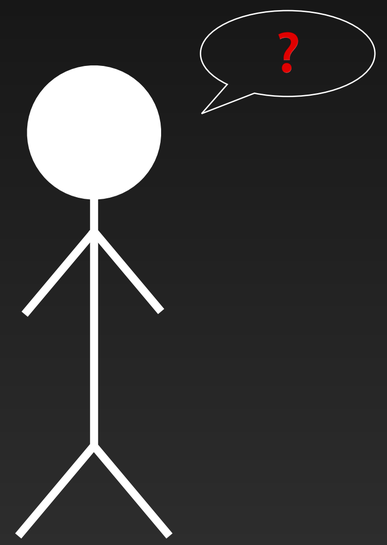
Trust

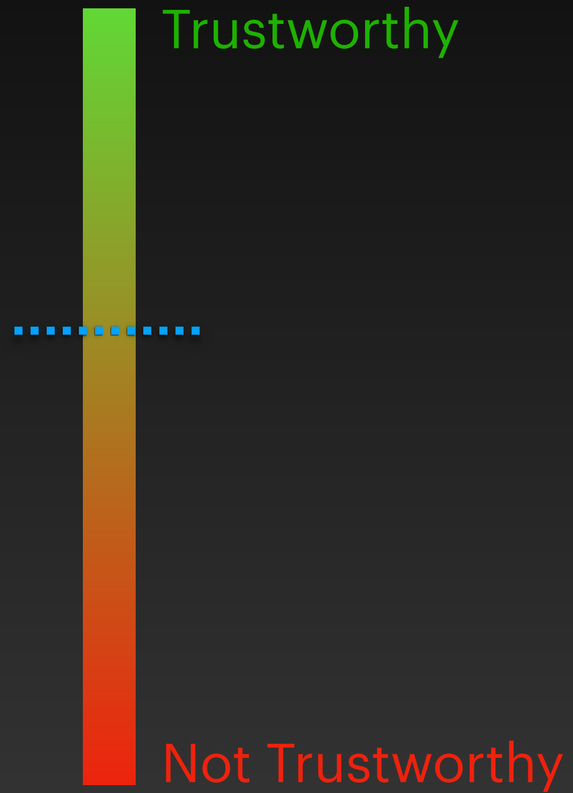


Trustworthy

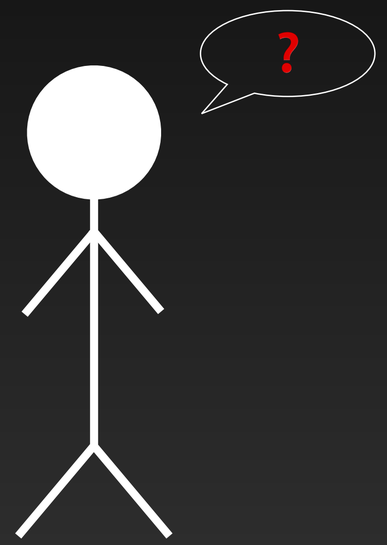
Not Trustworthy

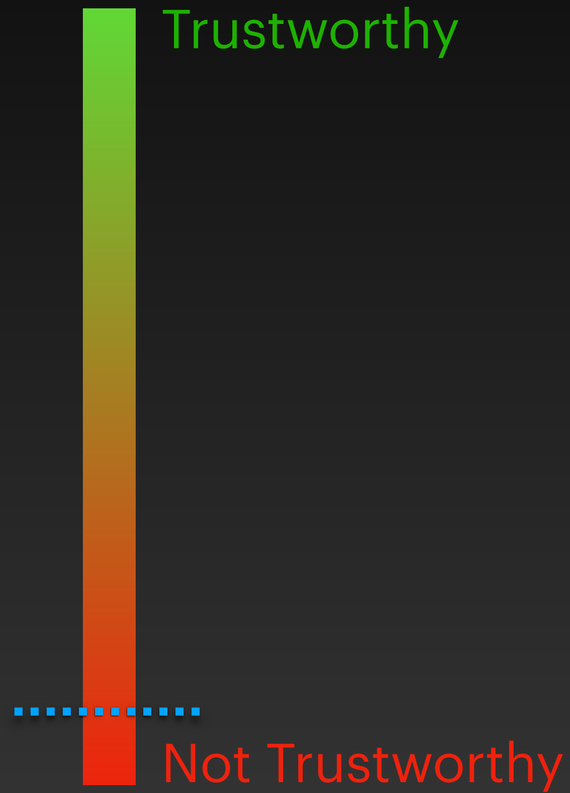
Trust



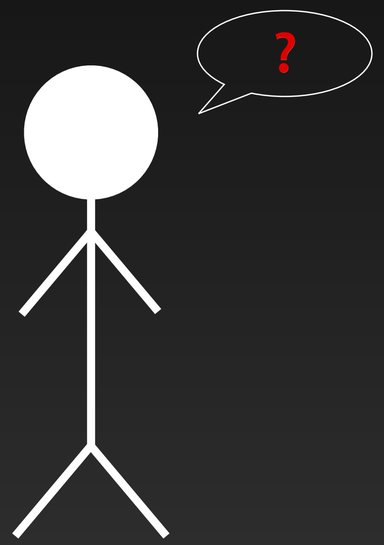


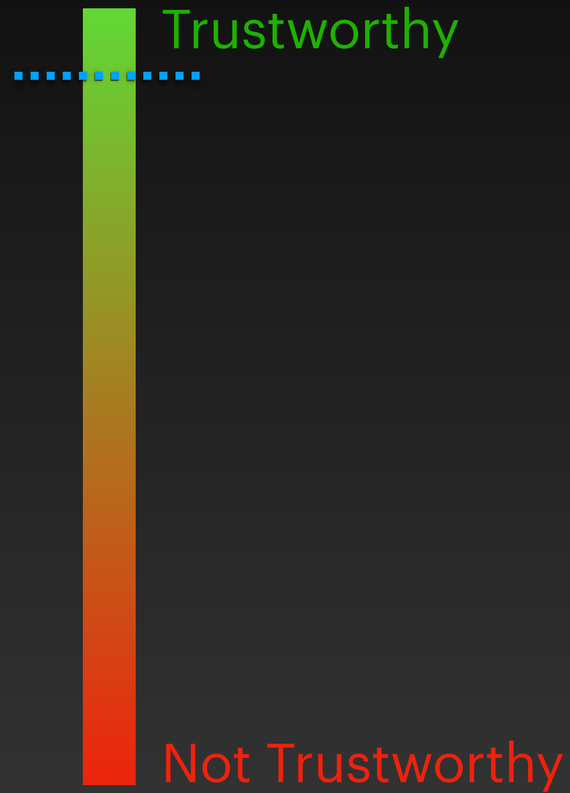
Trust



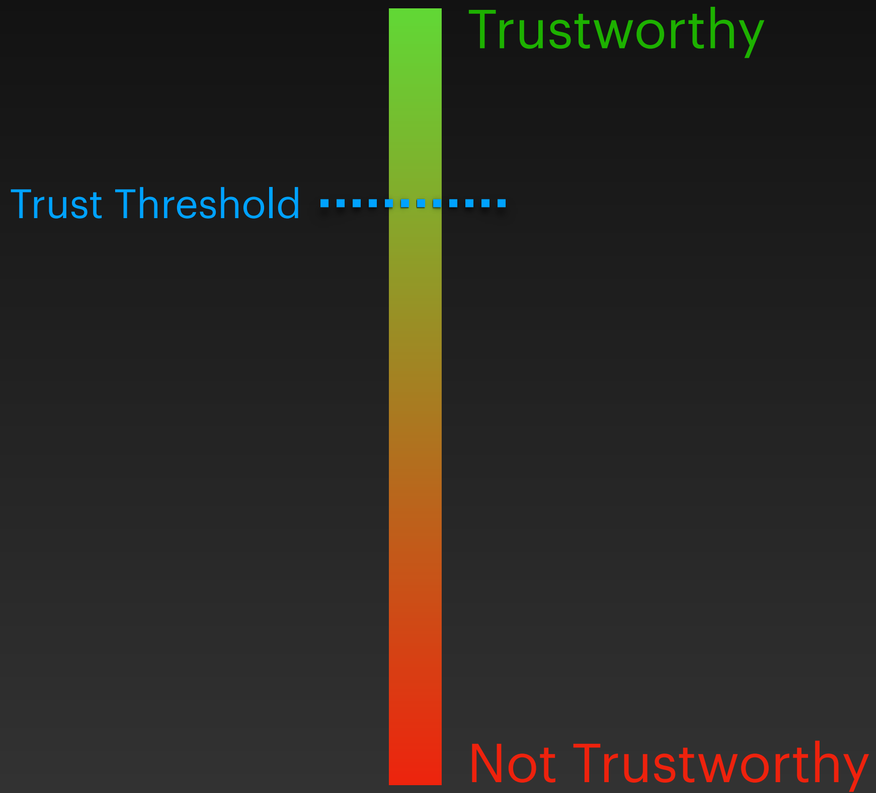


Trust





Trust



Trust Games

Trustor

Trust Games

Trustor

Trustee

Trust Games

Trustor

✅ Trustworthy

Trustee

Trust Games

Trustor

✅ Trustworthy

✅ Trusting

Trustee

Trust Games

Trustor

✅ Trustworthy

✅ Trusting

Trustee

❌ Trustworthy

Trust Games

Trustor

✅ Trustworthy

✅ Trusting

Trustee

❌ Trustworthy

❌ Trusting → ✅

Trust Games

1,000,000

Trust Games

Trustor

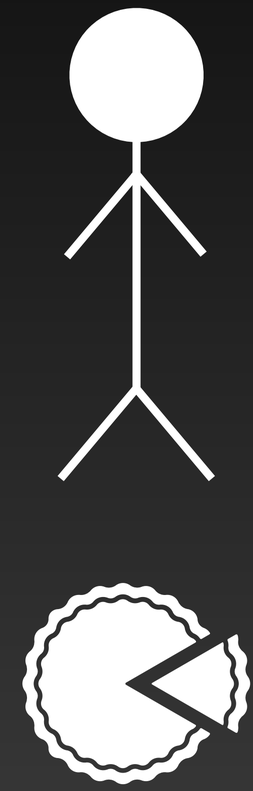
Trustee

V

Trust Games

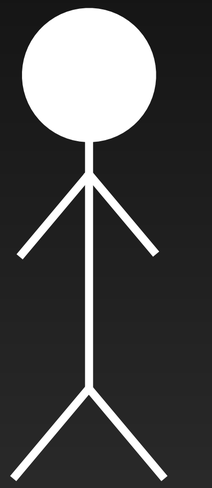
Trustor

Trustee

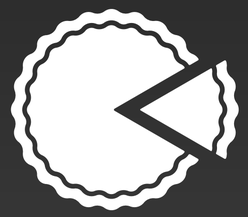


Trust Games

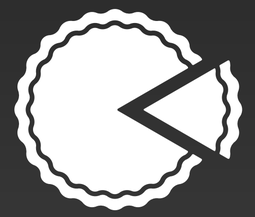
Trustor



$R_u = pV$



Trustee

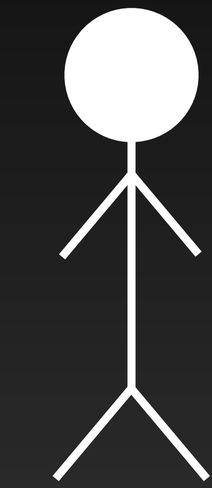


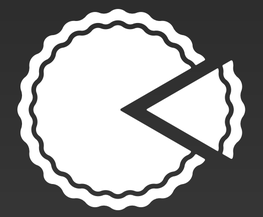
Trust Games

Trustor

Trustee

$$R_u = pV$$





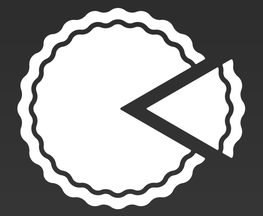
Trust Games

K

Trustor

Trustee

$R_u = pV$

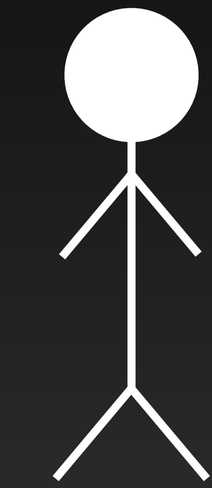


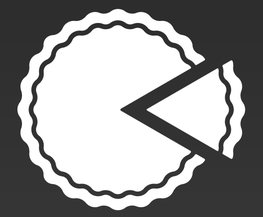
Trust Games

Trustor

Trustee

$$G_v = KR_u$$





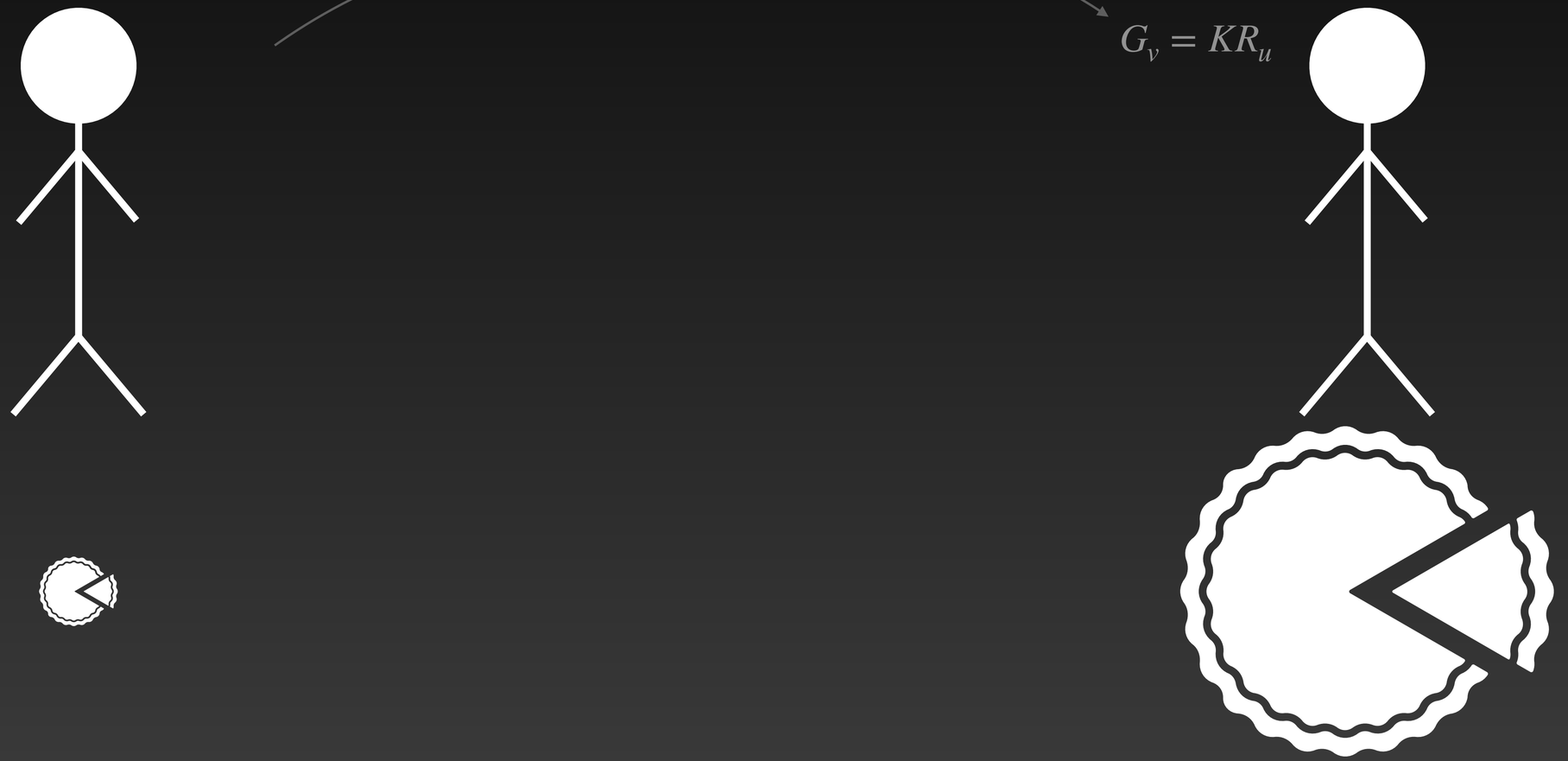
Trust Games

Trustor

Trustee

$R_u = pV$

$G_v = KR_u$



Trust Games

q

Trustor

Trustee

$R_u = pV$

$G_v = KR_u$

Trust Games

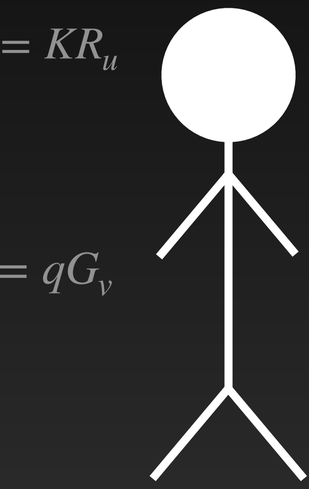
Trustor

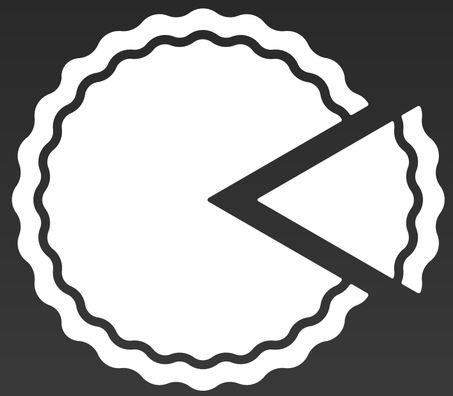
Trustee

$R_u = pV$

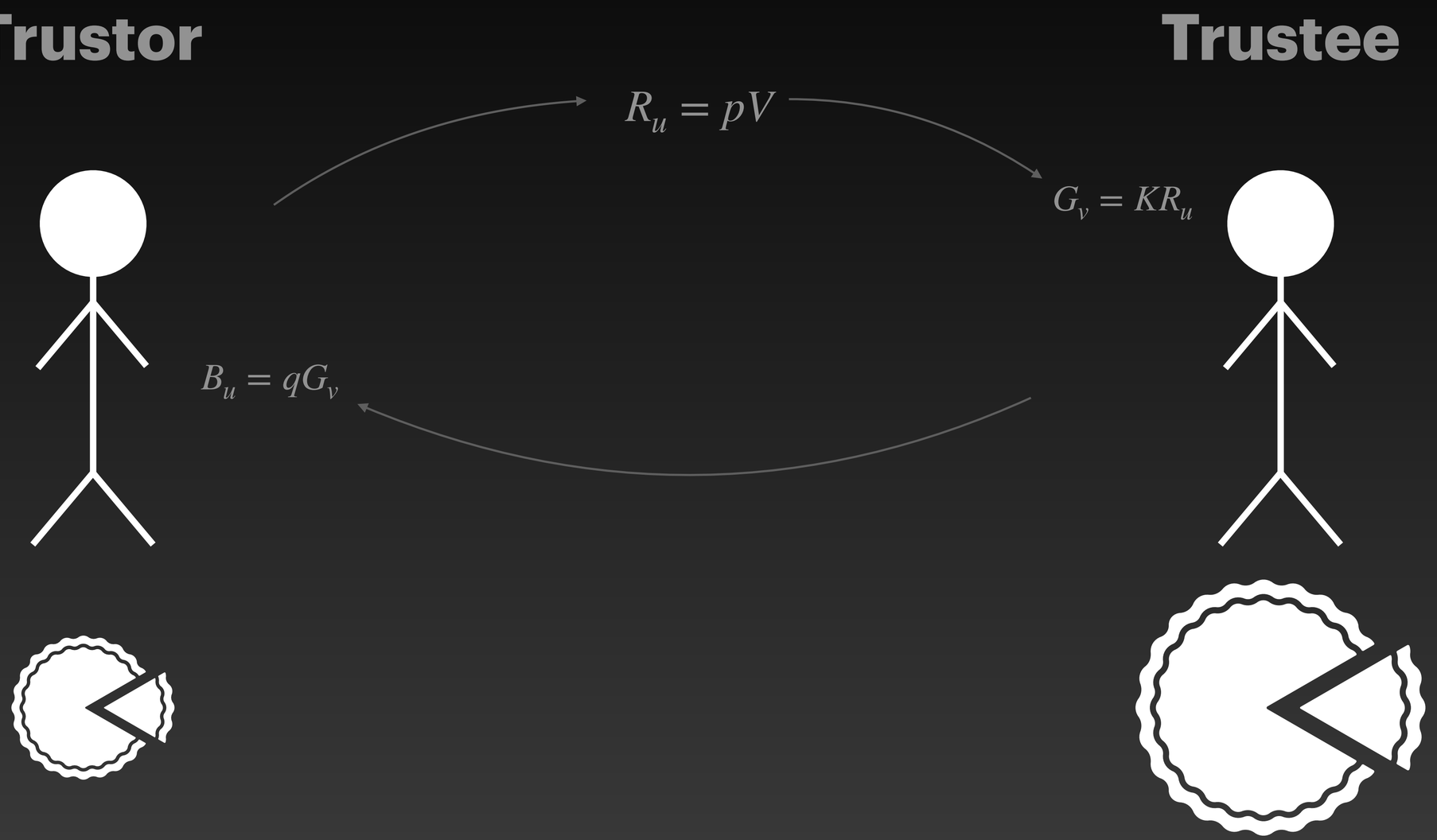
$G_v = KR_u$

$B_u = qG_v$

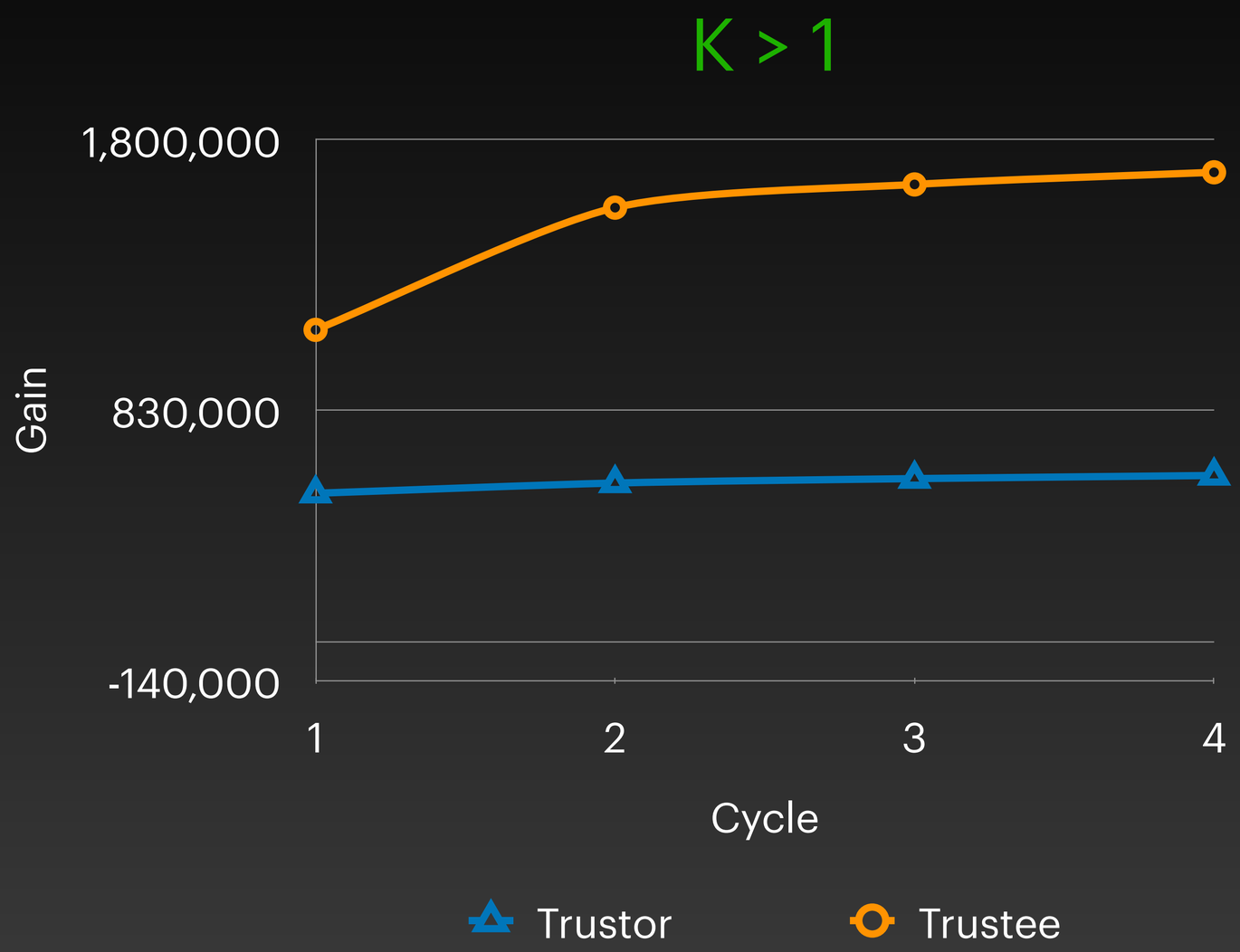




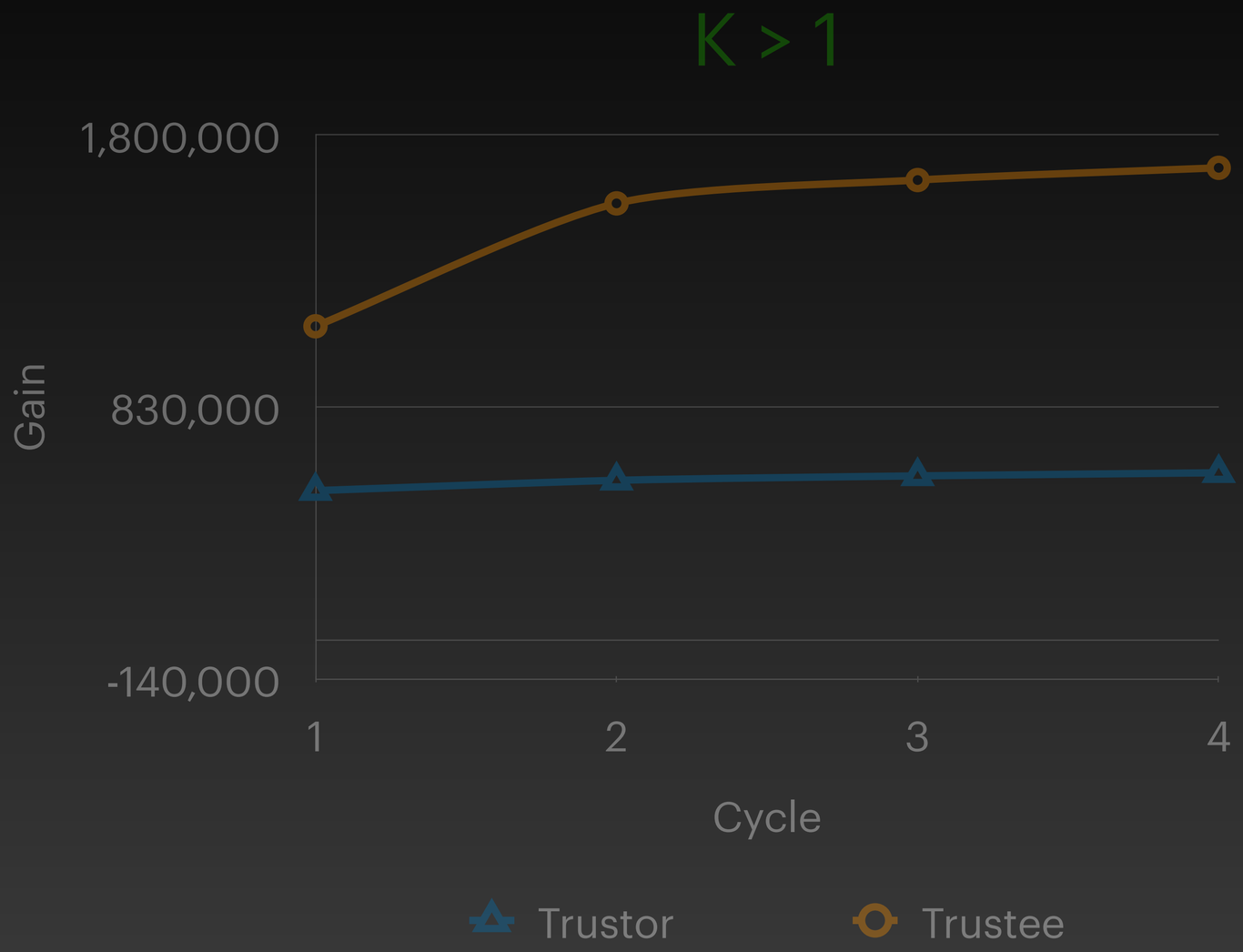
Trust Games

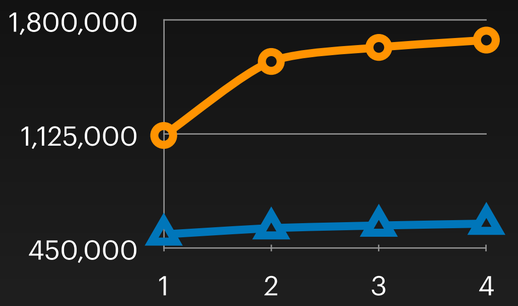


Simulation 1 – Adding Value

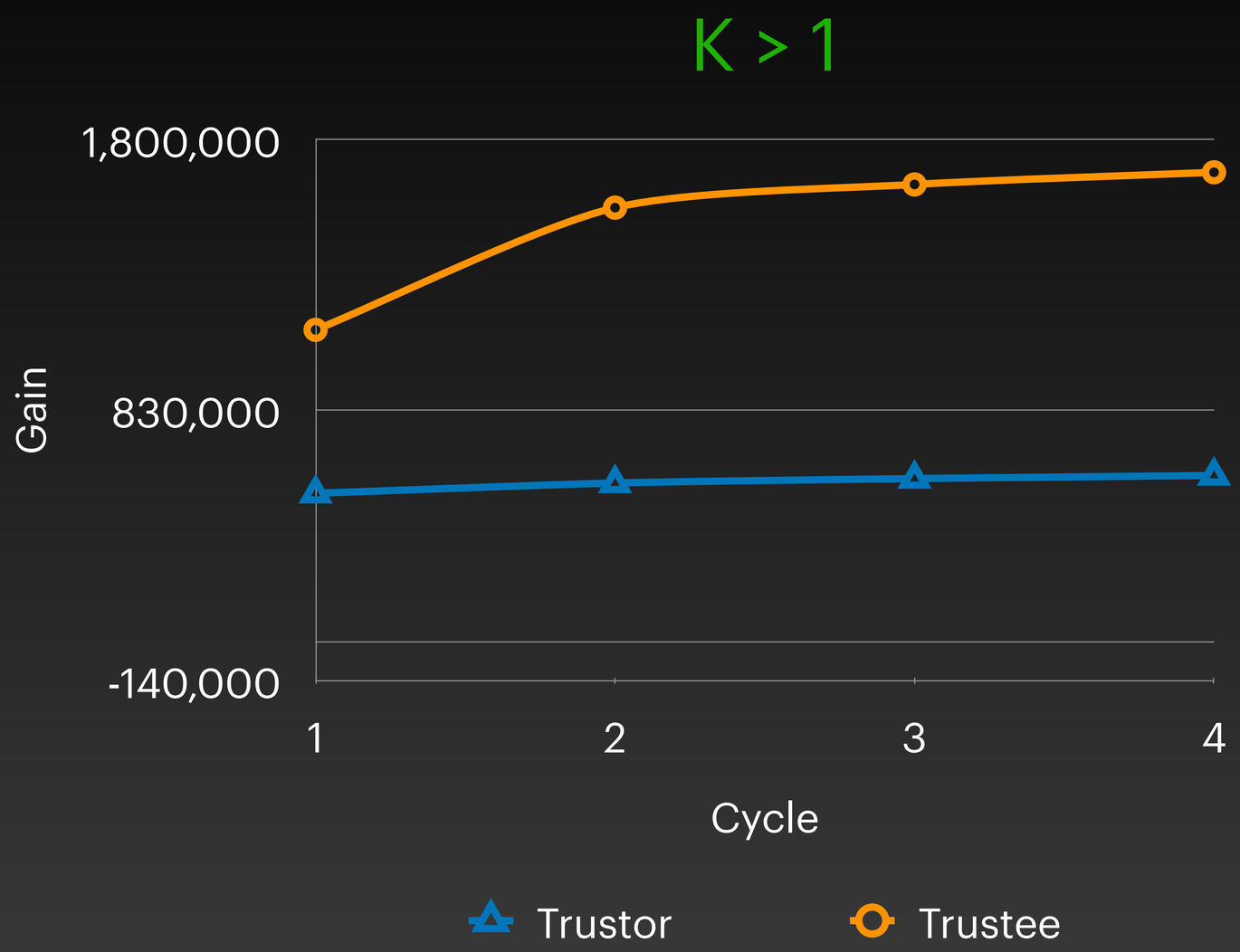


Simulation 1 – Adding Value

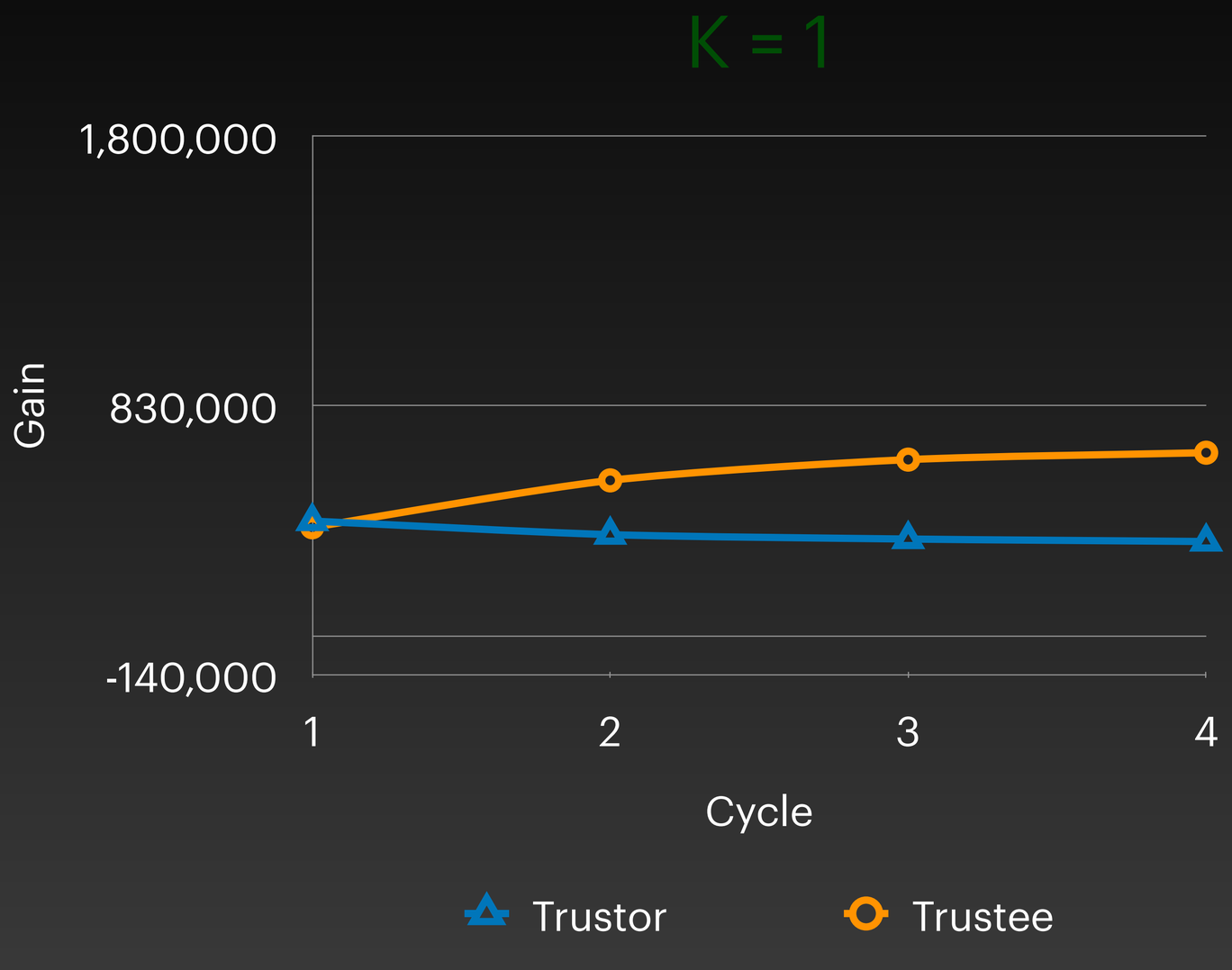




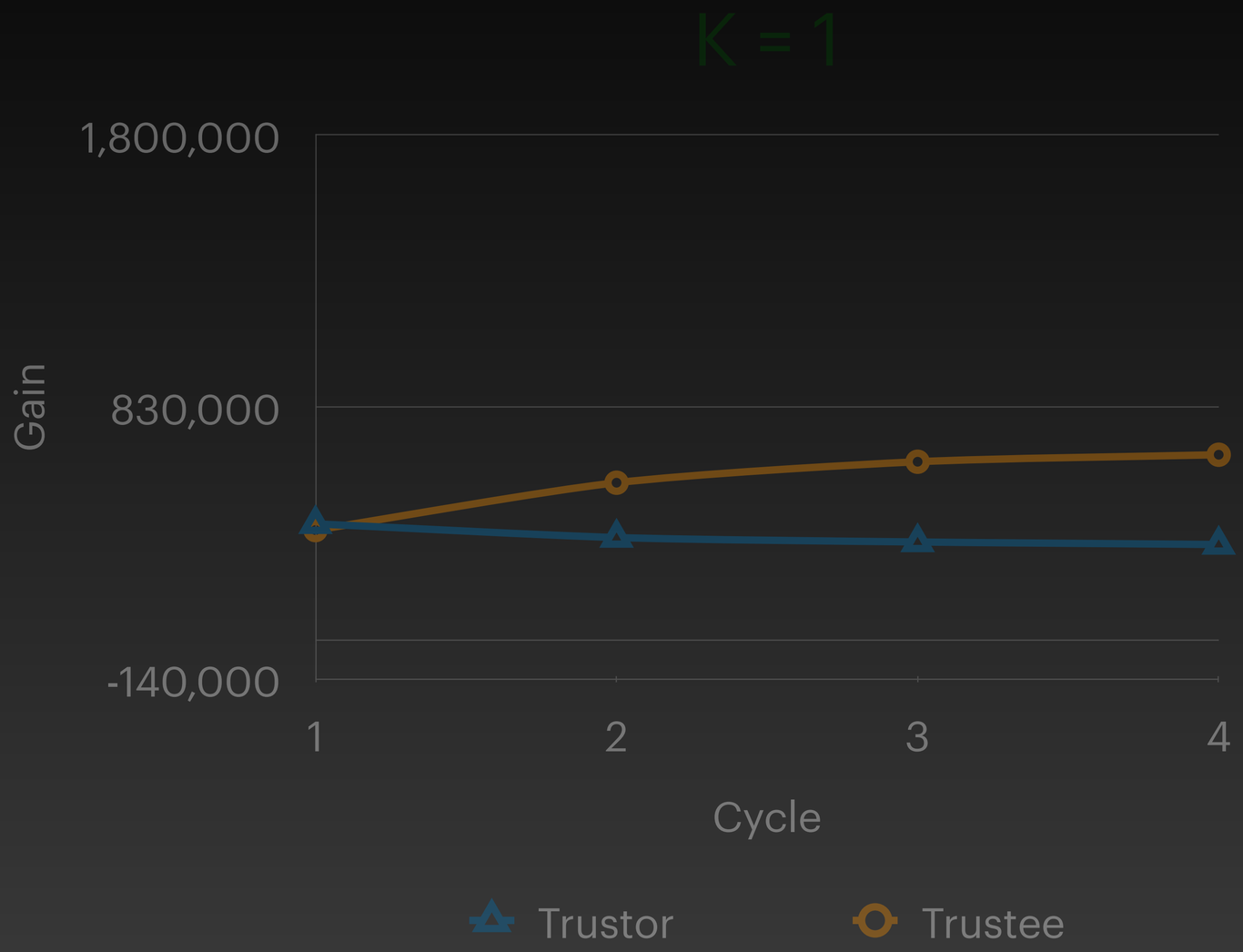
Simulation 1 – Adding Value

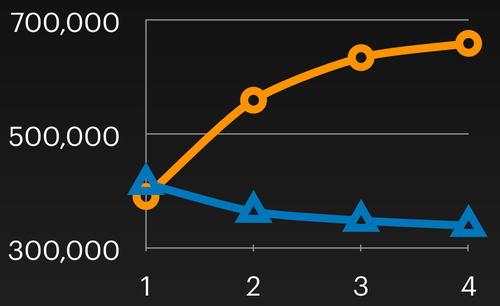


Simulation 2 – Neutral

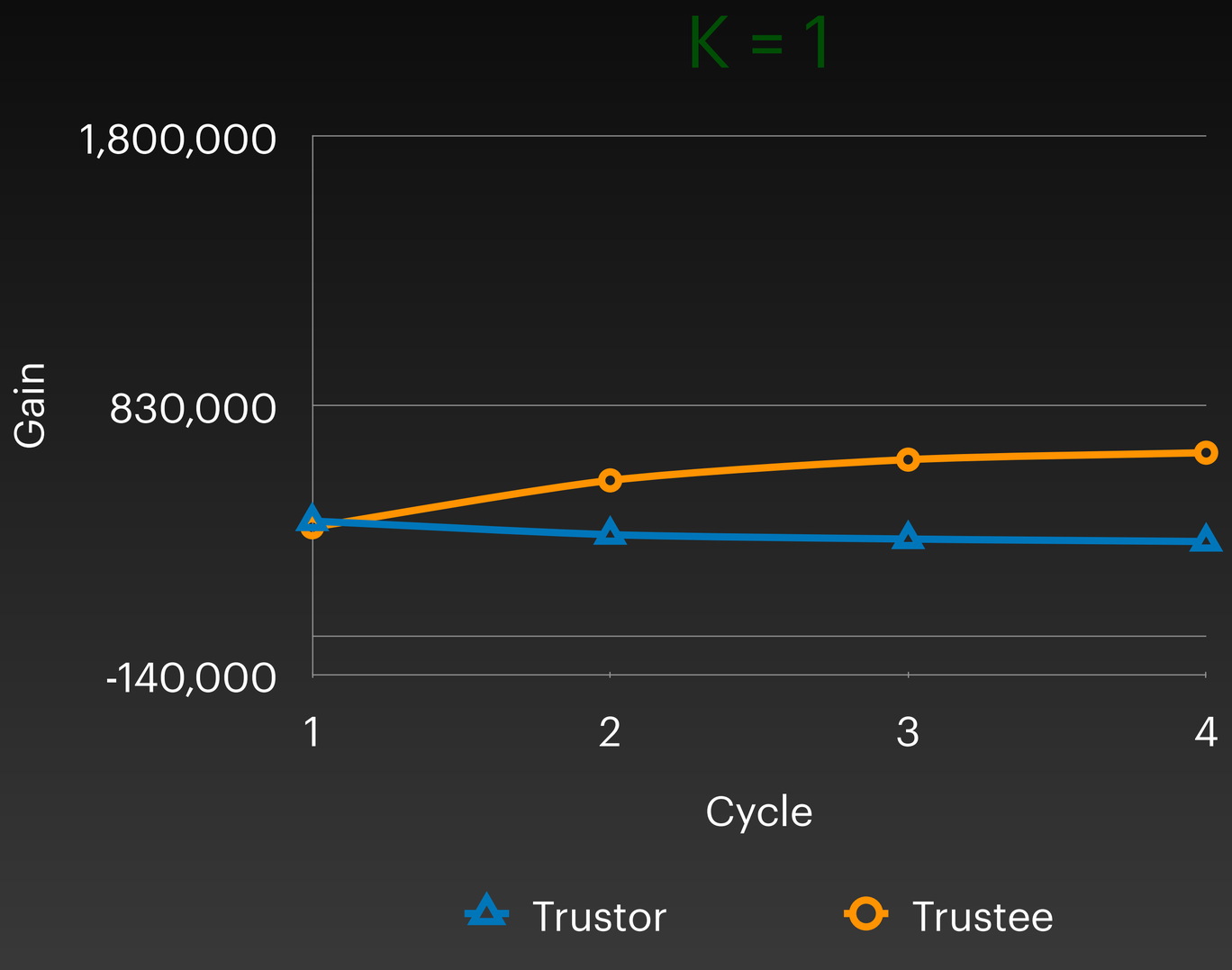


Simulation 2 – Neutral



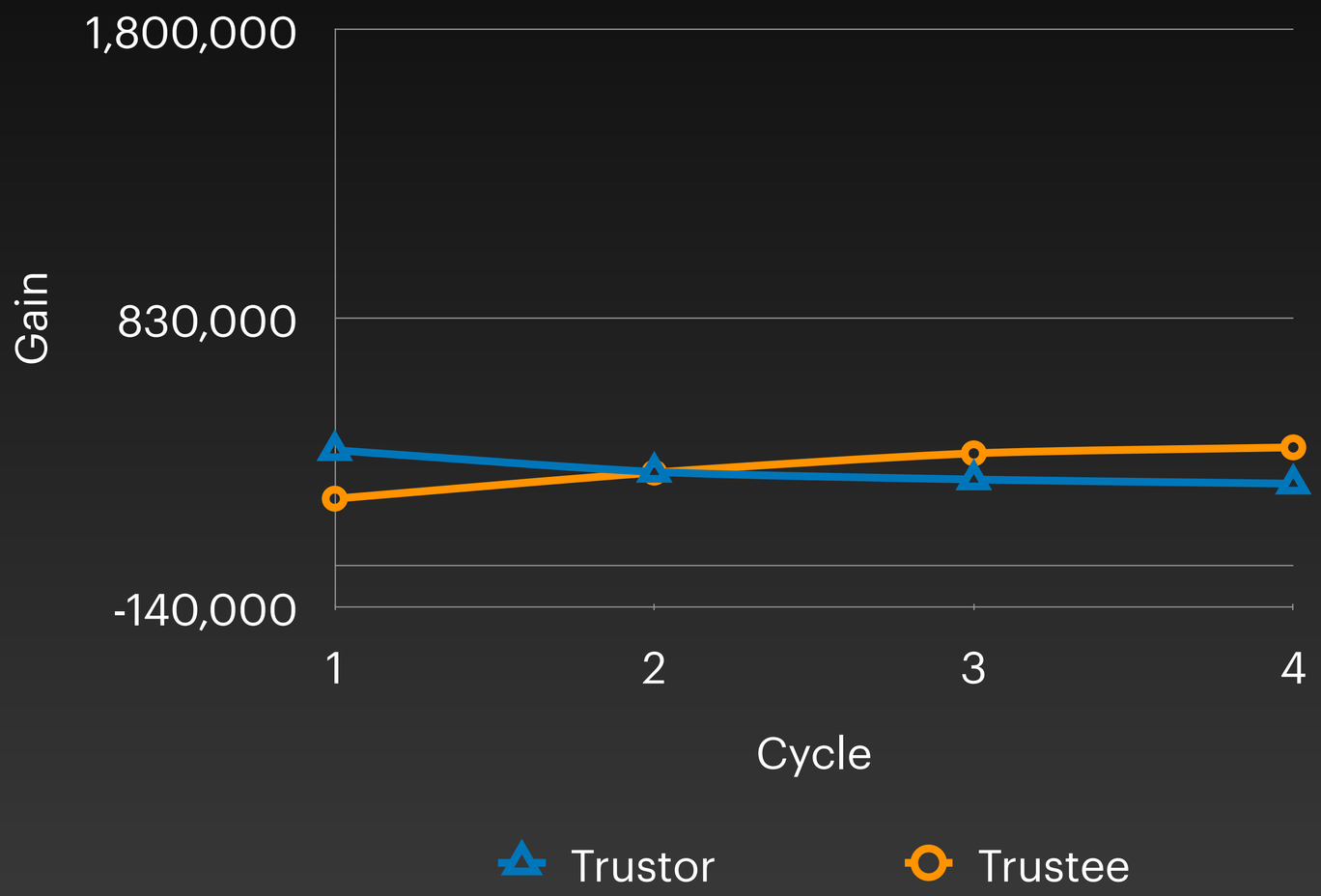


Simulation 2 – Neutral



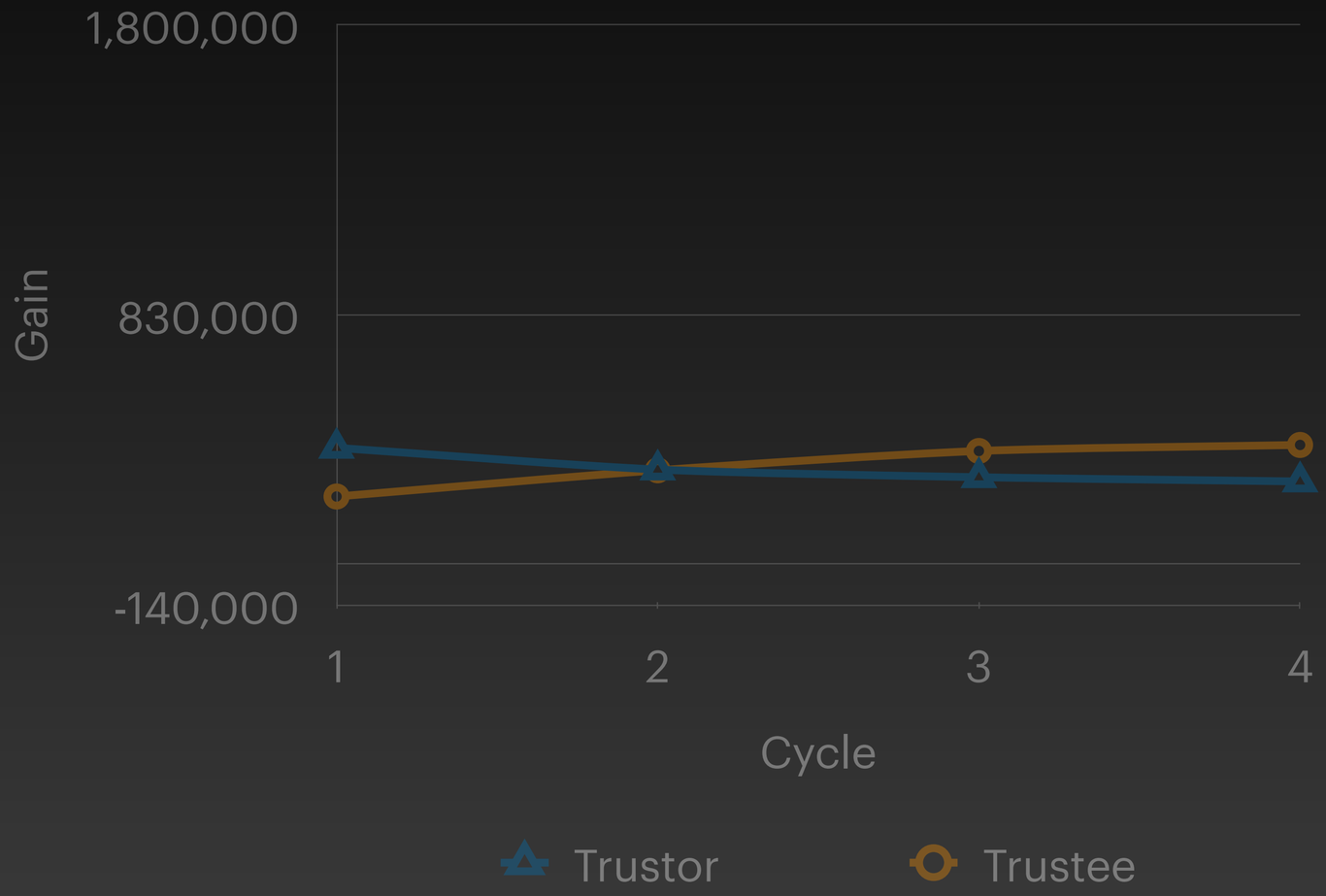
Simulation 3 – Causing Inefficiencies

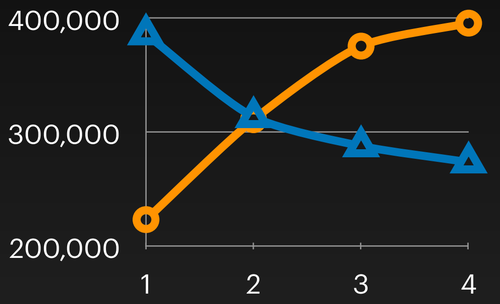
$0 \leq K < 1$



Simulation 3 – Causing Inefficiencies

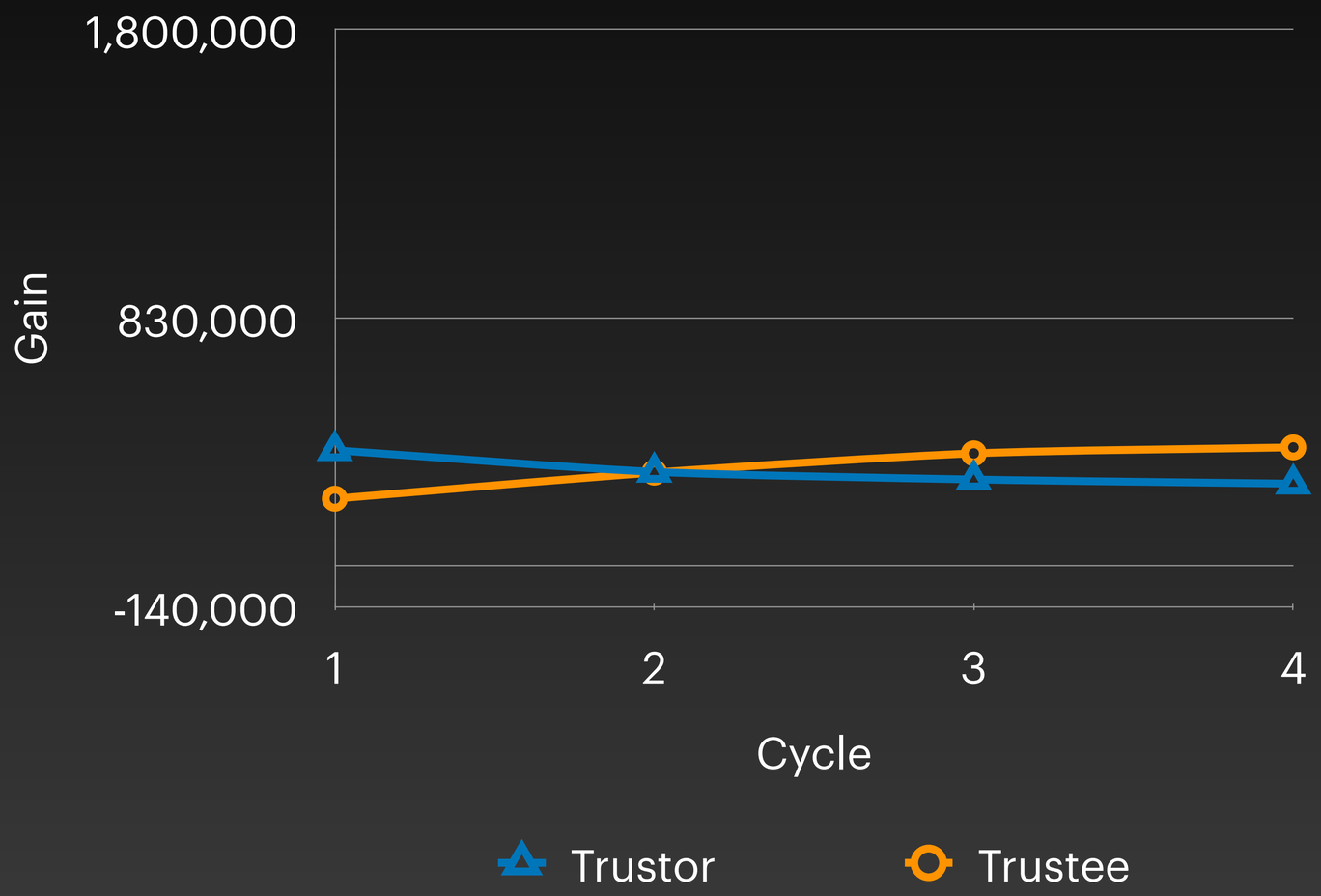
$0 \leq K < 1$



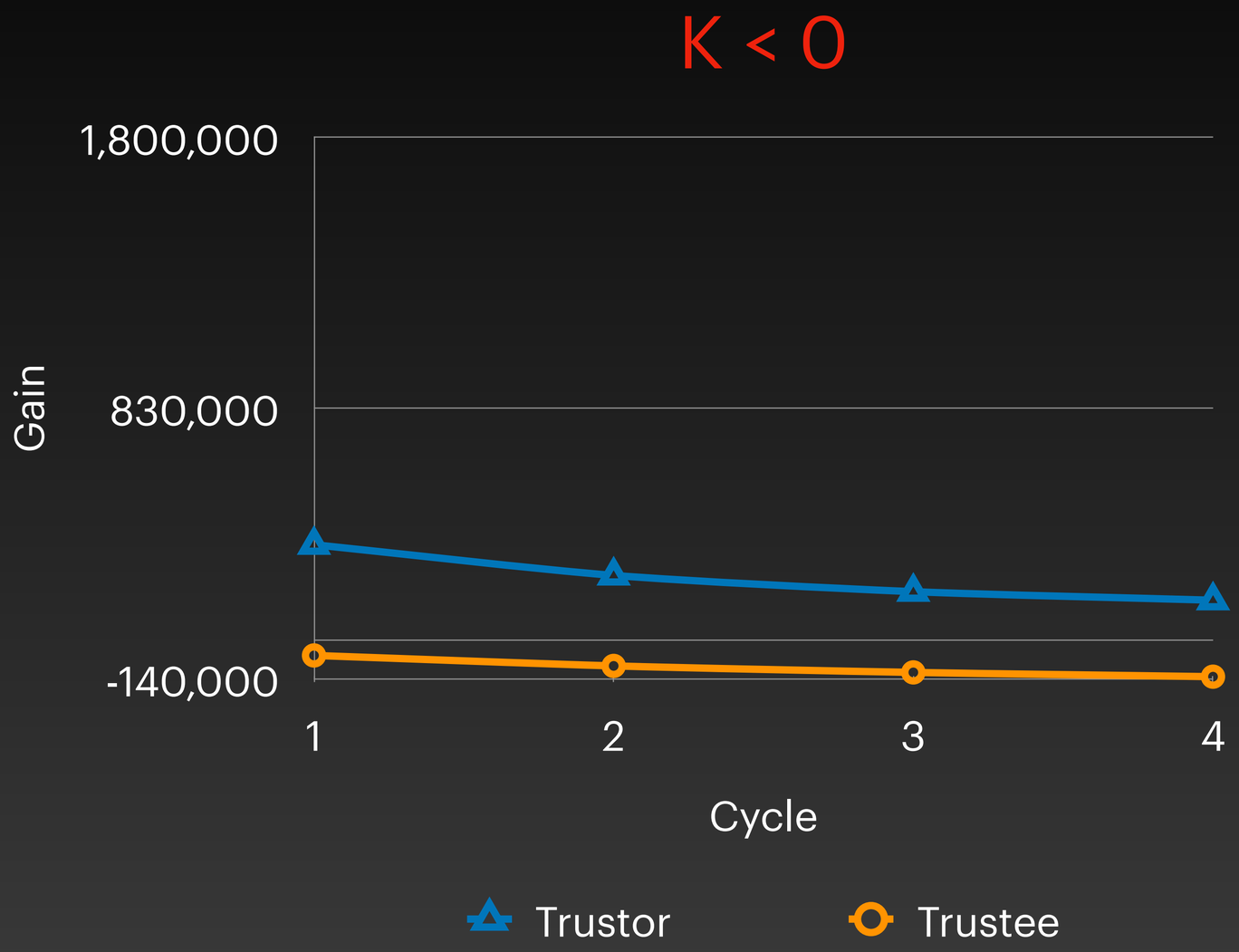


Simulation 3 – Causing Inefficiencies

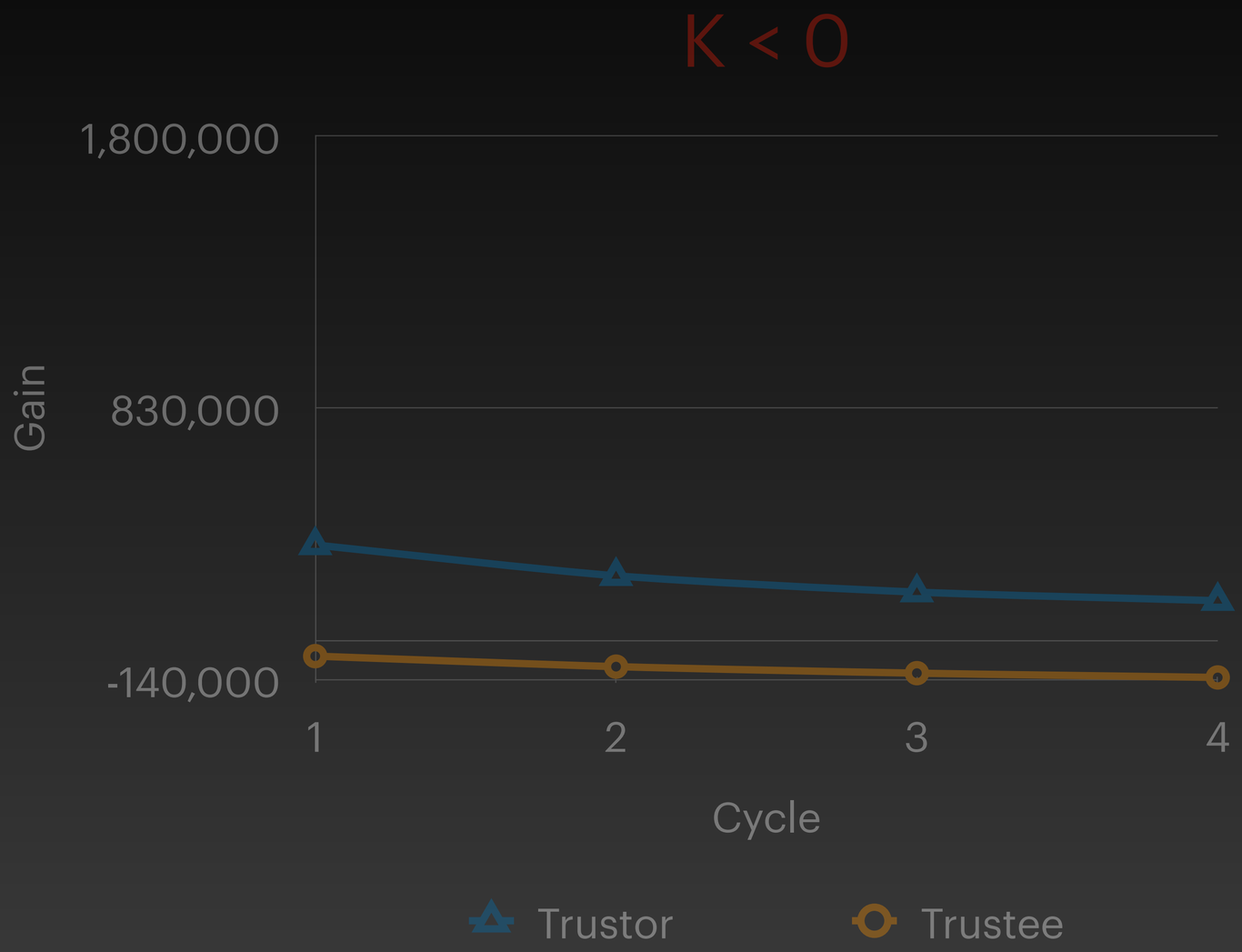
$0 \leq K < 1$

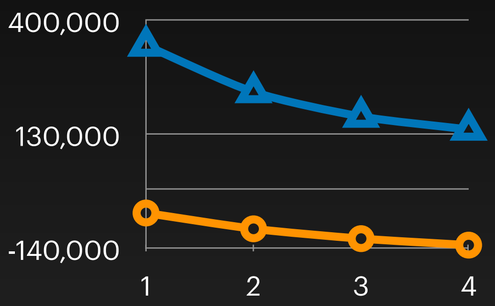


Simulation 4 – Rapid Erosion of Trust

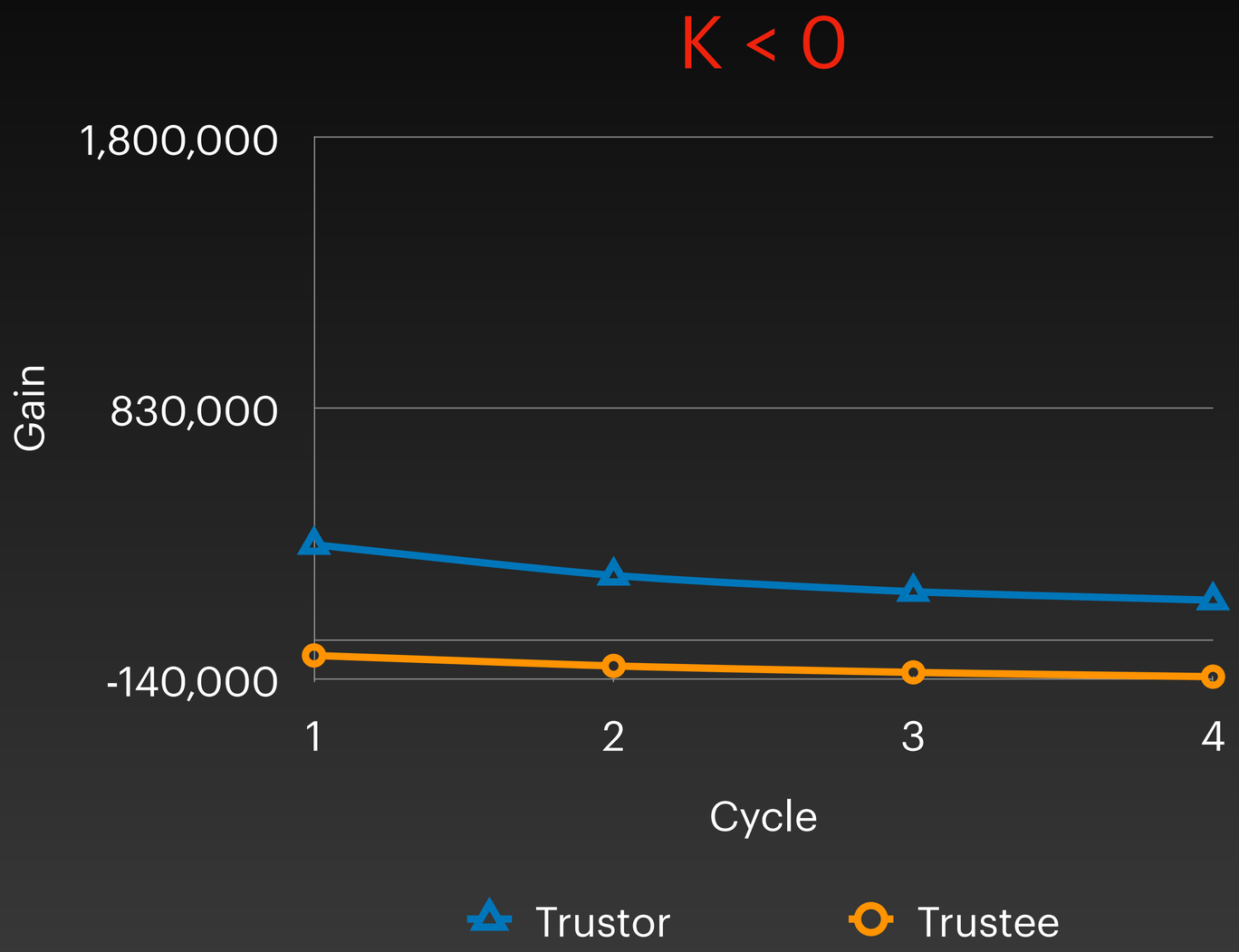


Simulation 4 – Rapid Erosion of Trust

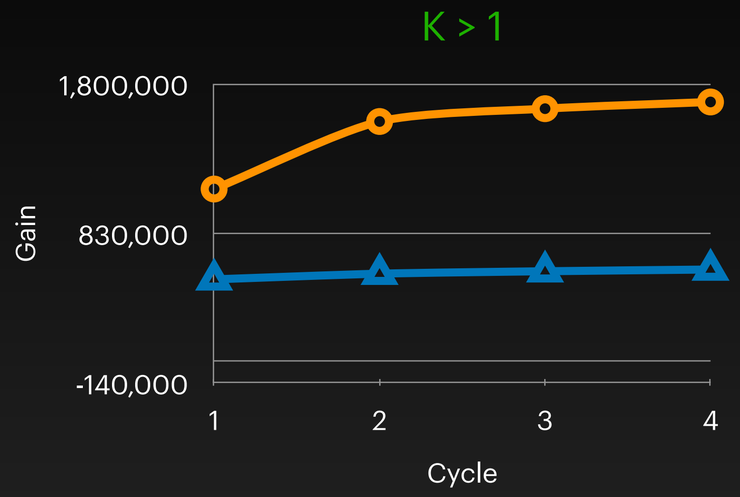




Simulation 4 – Rapid Erosion of Trust

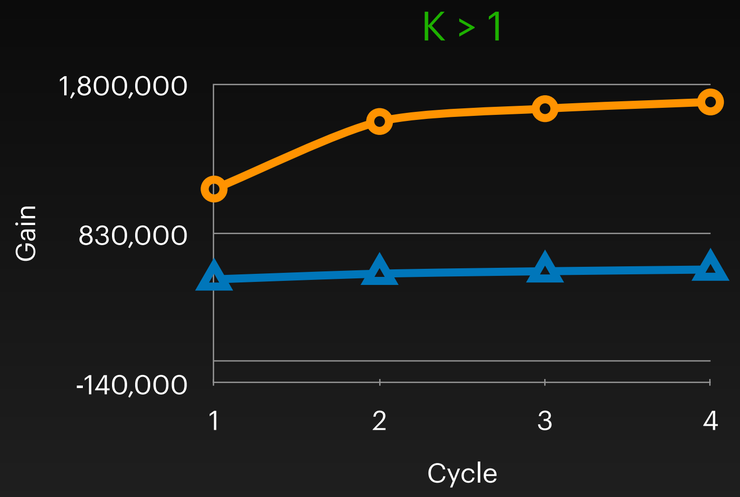


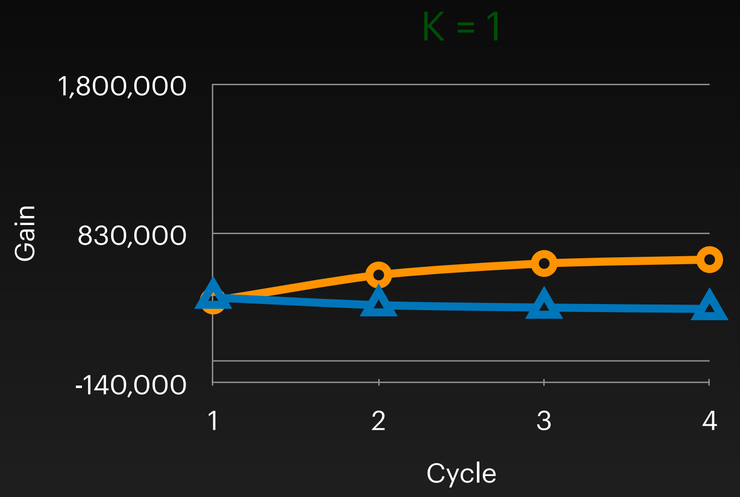
Simulations Side by Side

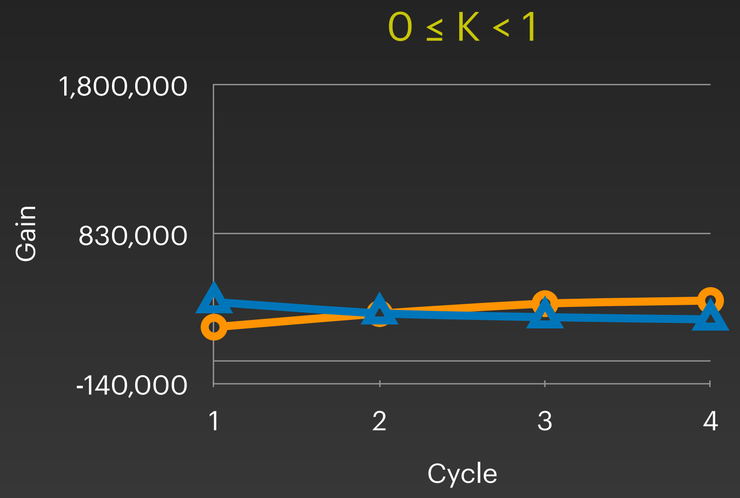


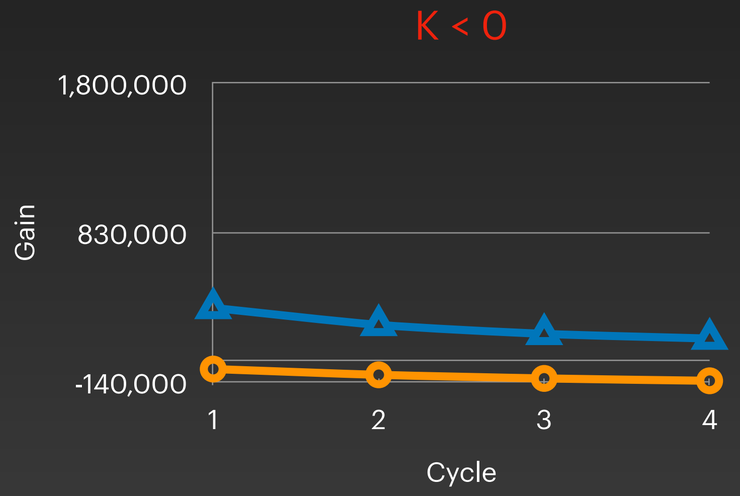
Trustor Trustee

Simulations Side by Side



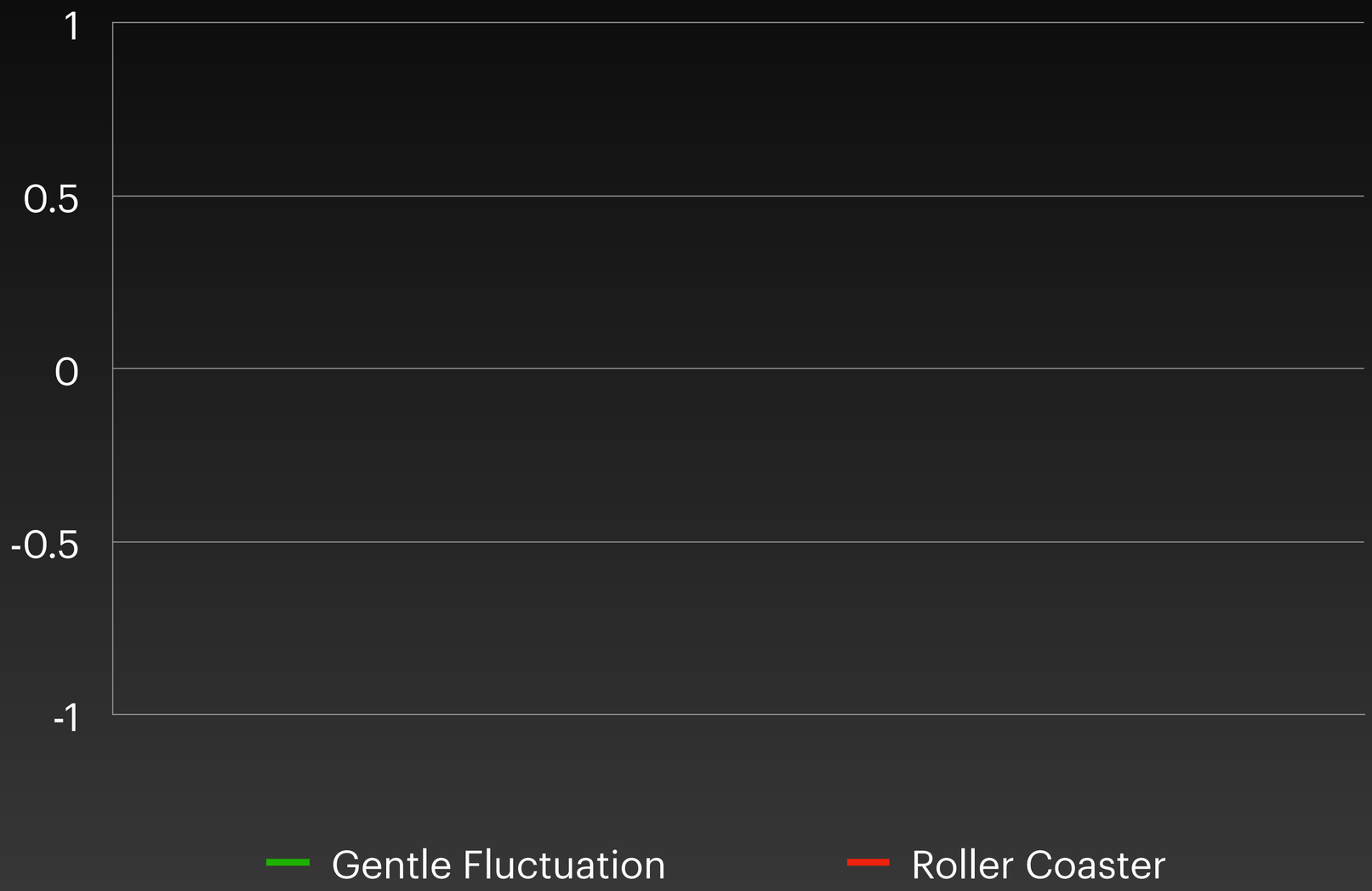




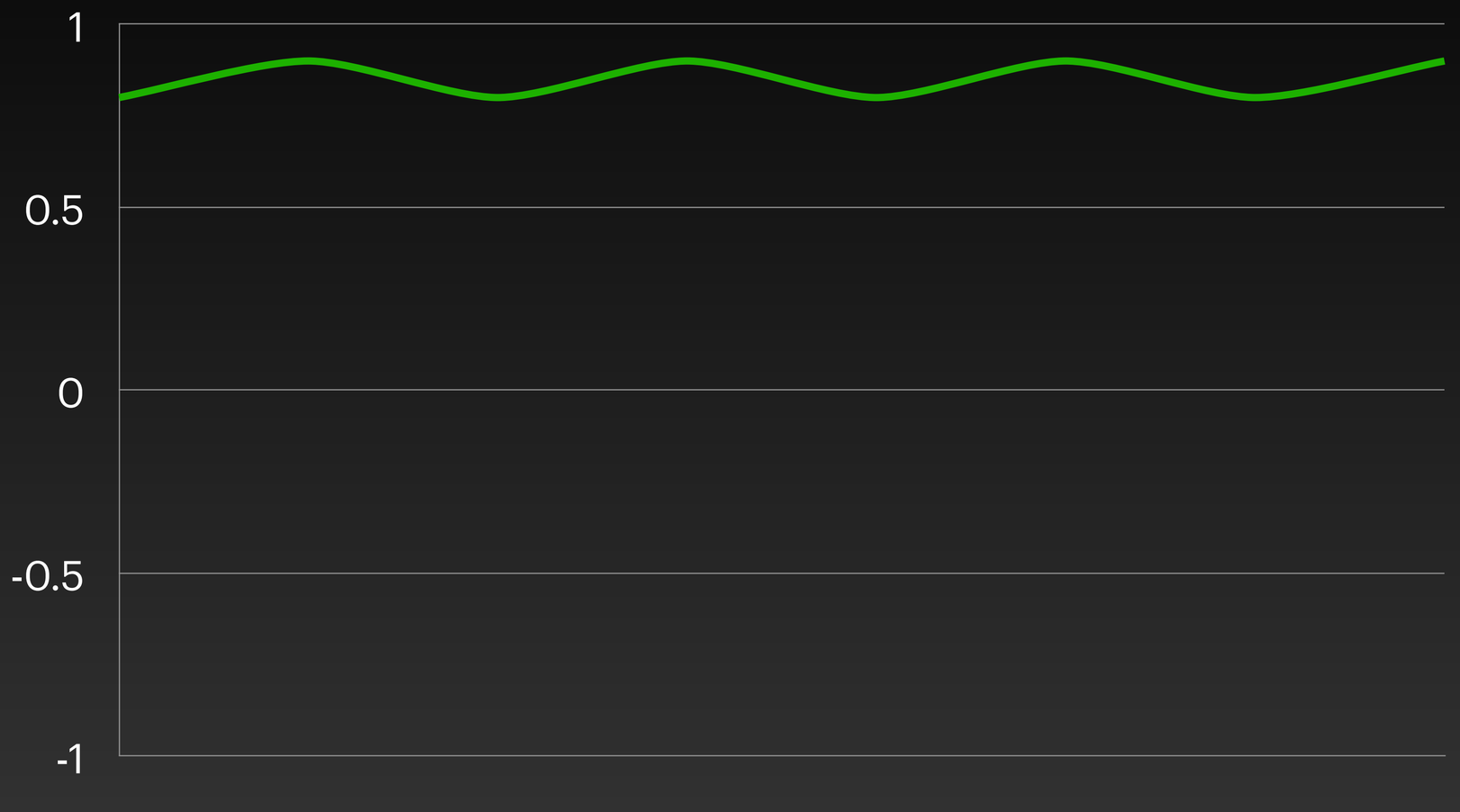


Trustor Trustee

Trust Over Time



Trust Over Time



Trust Over Time

Let’s Talk. Here and Online.

Dalmo Cirne
@d_cirne
https://www.linkedin.com/in/dalmocirne/