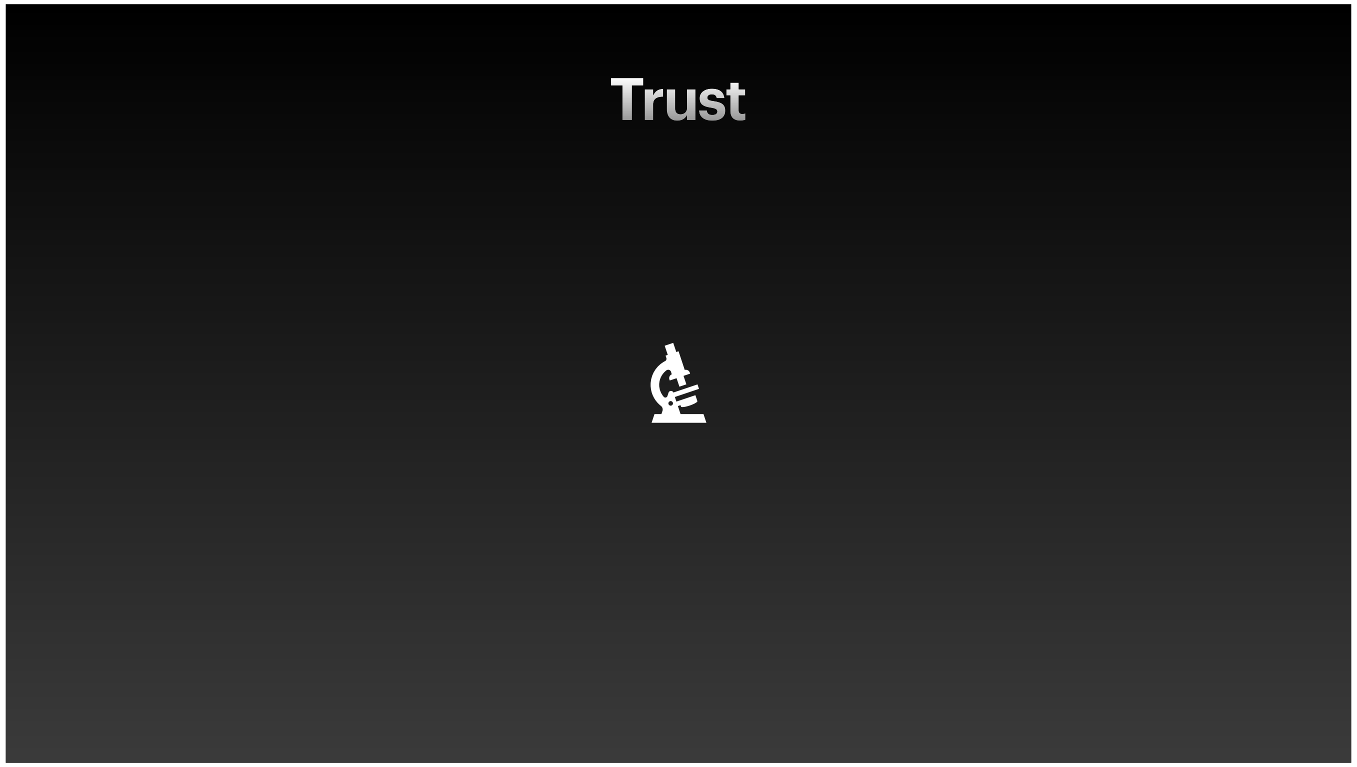
# Fostering Trust and Quantifying Value of AI and ML

Dalmo Cirne
https://dalmocirne.com

Veena Calambur
https://www.linkedin.com/in/veena-calambur/

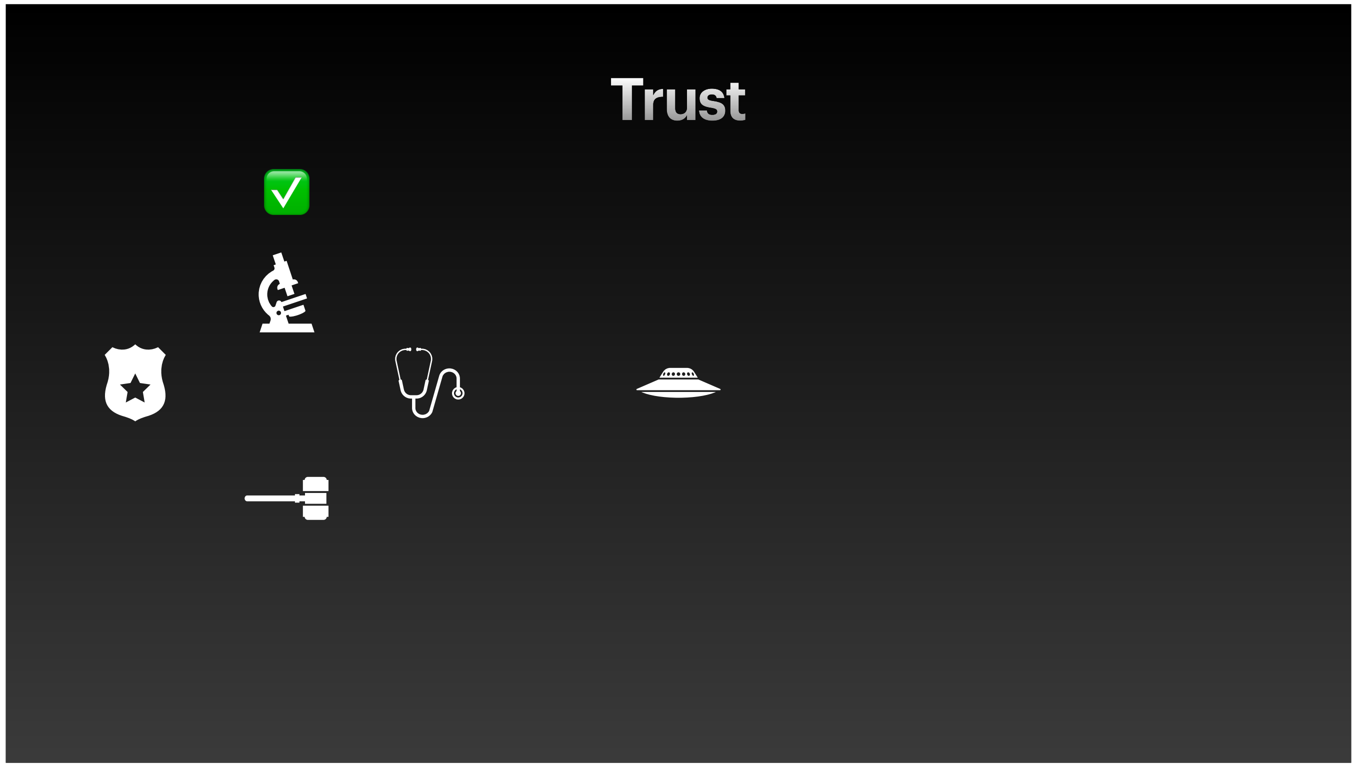
IARIA 2024



## Trust

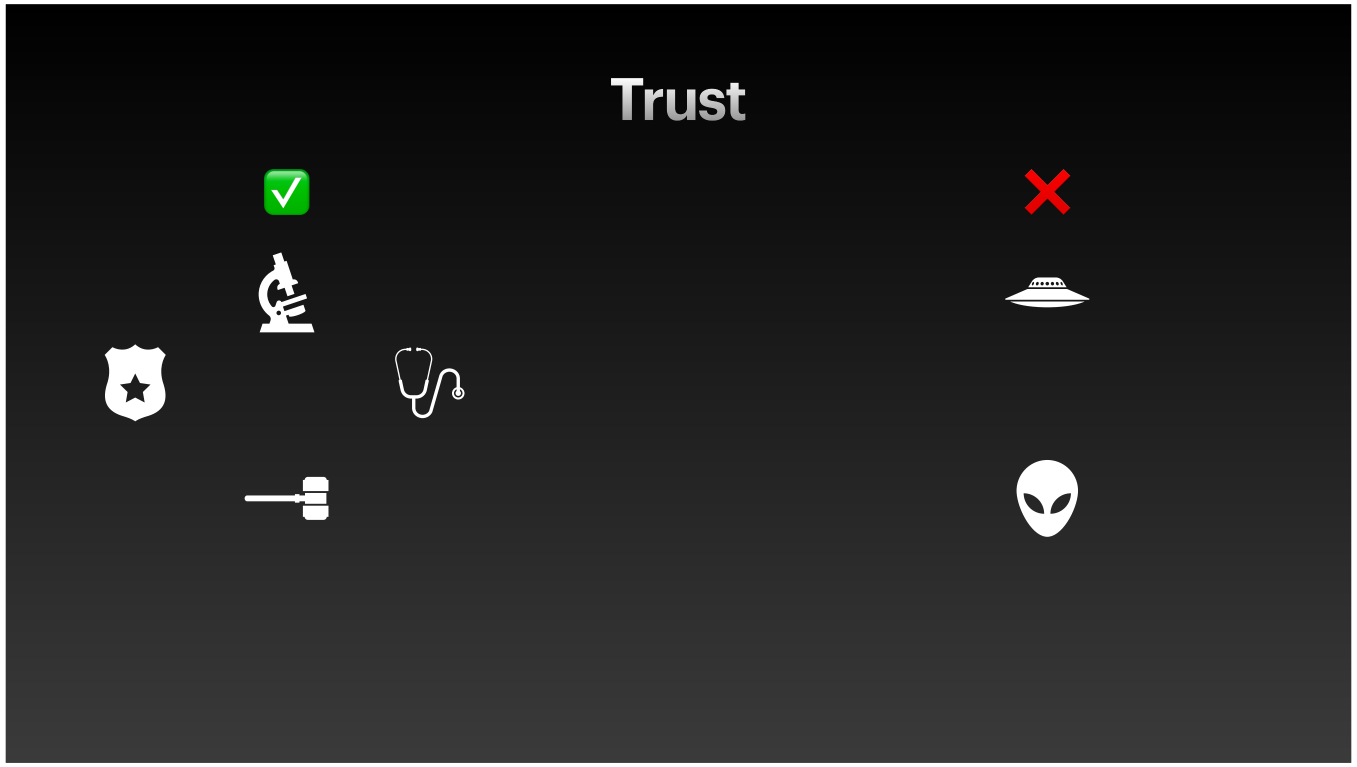
🔬

- How do we go about building trust? In western societies, we generally trust scientists, police officers, doctors, and judges. Here we are not referring to blind trust, but an inclination to trust first, but that trust can be lost (for various reasons).



## Trust

- But how about aliens on a spaceship? I would argue that our first reactions would not be of trust (we don’t know their intent, their military capabilities, and so on)
- That said, it is possible that they can build trust with us



# Trust

✓

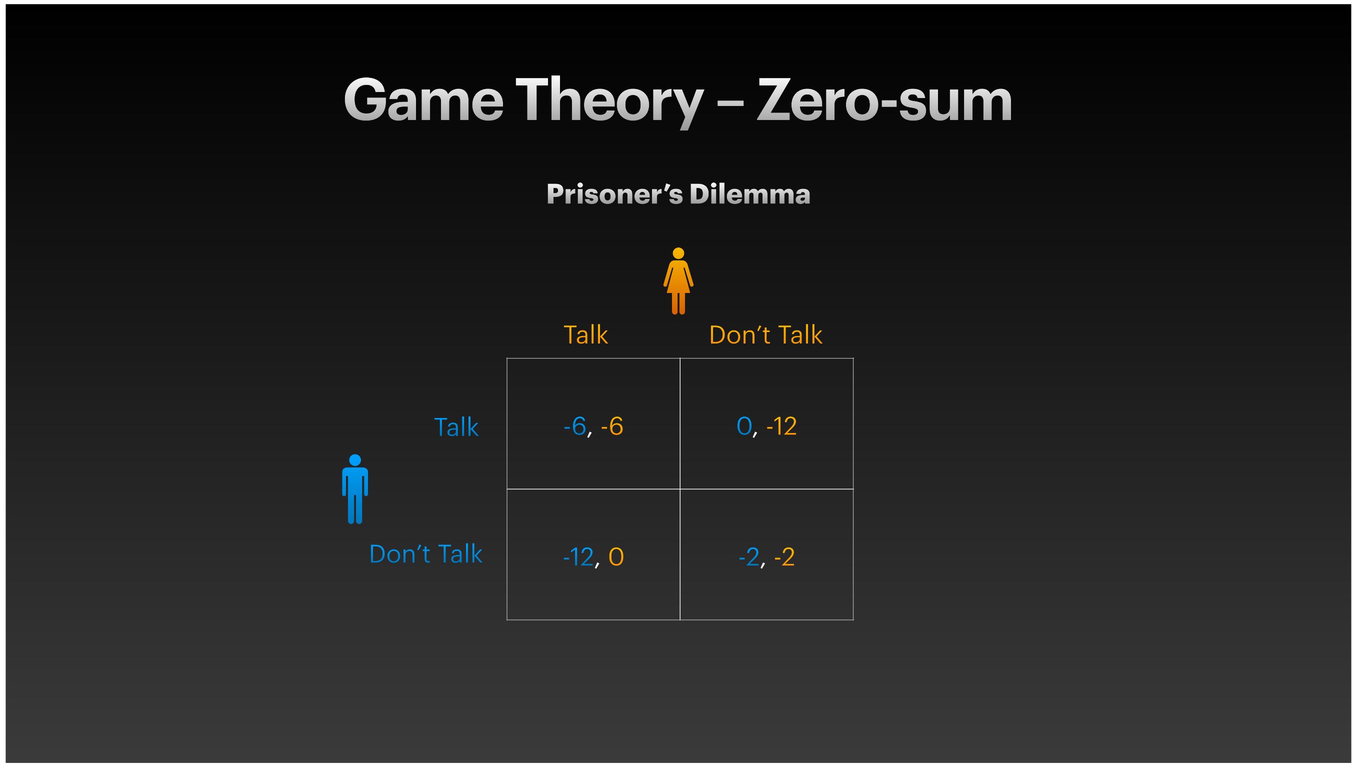
✗

## Qualitative

- Much is said about responsible AI/ML. Many claim to be bastions of responsibility
- Is it enough to say that your institution adopts “best practices”? What even are those practices?
- There are several cases throughout history when some proclaimed to have people’s best interests, yet the results were incredibly damaging. A few well known examples are: FTX, Theranos, Bernie Madoff, Volkswagen, Wirecard

# Qualitative vs Quantitative

- Besides taking people’s words for it, we currently lack a quantitative of assessing trust in AI/ML
- Quantification by itself would prevent fraud, but it would add significant friction and provide auditability

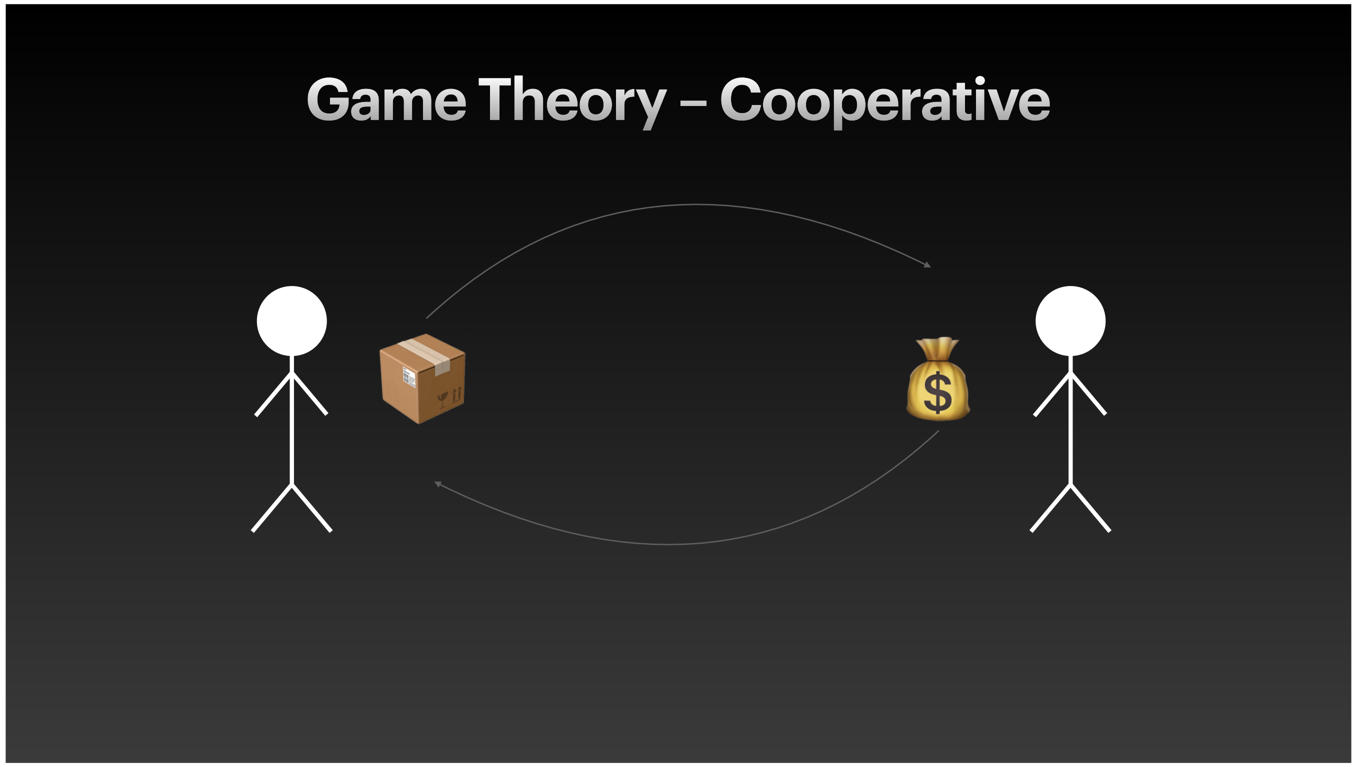


## Game Theory – Zero-sum

### Prisoner’s Dilemma

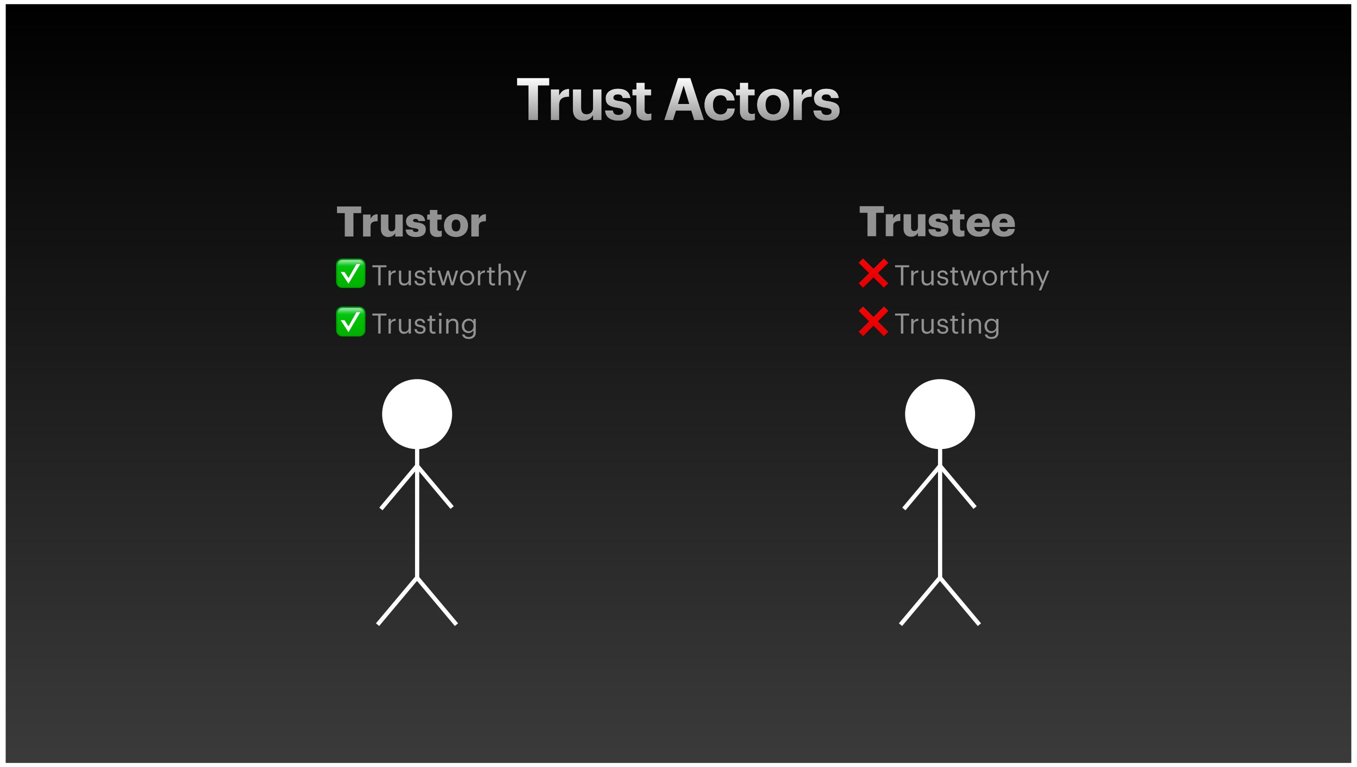
|  | Talk | Don’t Talk |
|---|---|---|
| Talk | -6, -6 | 0, -12 |
| Don’t Talk | -12, 0 | -2, -2 |

- Game theory is a mathematical framework for modeling and analyzing situations where decision-makers interact
- Imagine Bonnie and Clyde were captured and were interrogated separately. Although both stay silent and get a smaller sentence, officers could tell Bonnie that Clyde no longer likes her and will tell everything. Bonnie may betray her partner and start taking, trying to get immunity. Clyde may do the same. In the end the best strategy for them is to talk (Nash Equilibrium)
- In a zero-sum game for one to win the other has to lose



# Game Theory – Cooperative

- Zero-sum games often only involve one interaction among participants and incomplete information
- In cooperative games, participants interact multiple times and much more information about each other
- One way to start building trust is the fair trading of value. But someone has to take the first step. Here we argue that the first step must be taken by the party building and selling the product/service. They had to invest in building it before a relationship of trust was formed



## Trust Actors

**Trustor**

✅ Trustworthy

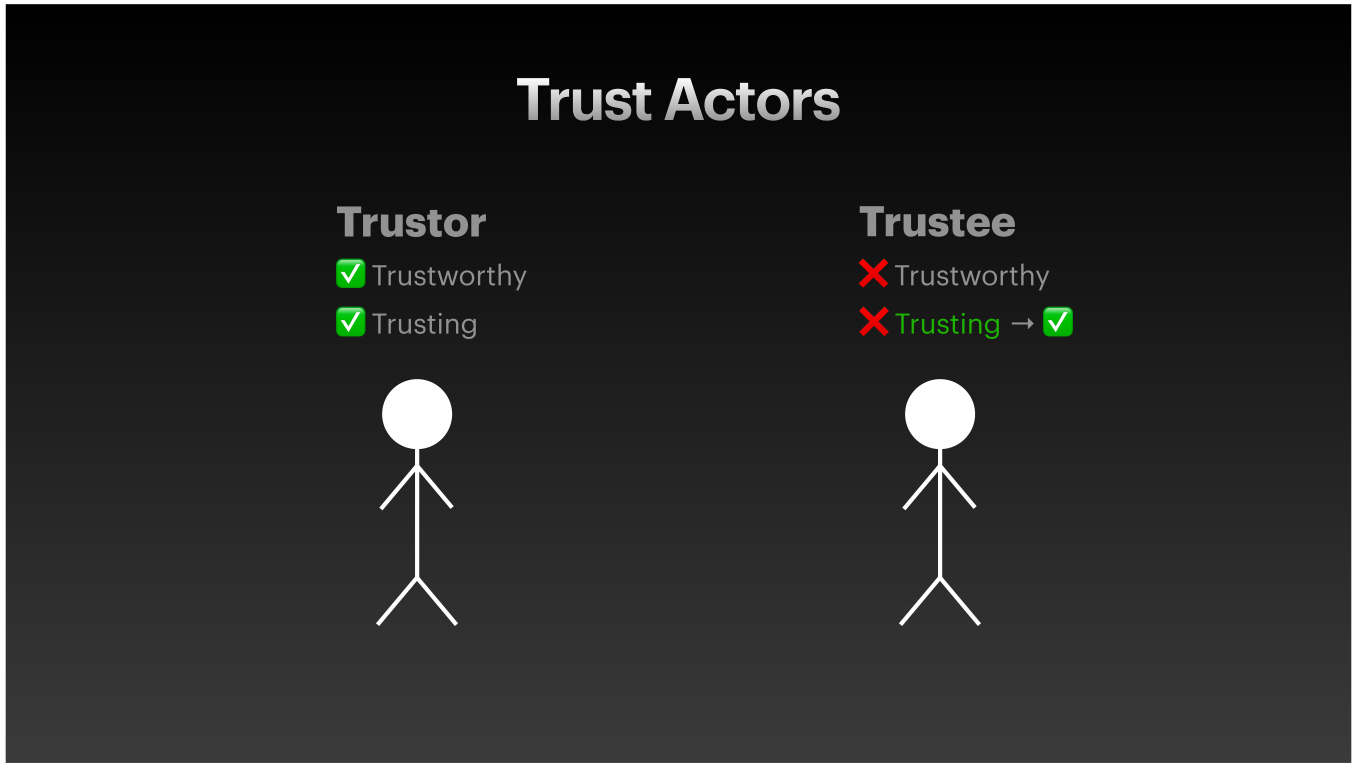
✅ Trusting

**Trustee**

❌ Trustworthy

❌ Trusting

- There are at least two actors in cooperative games
- Trustors must be trustworthy and trusting
- Trustees are not required to be trustworthy (imagine they don’t like the same sports team as you, surely they are not trustworthy), yet you can collaborate
- They don’t have to trust you either



## Trust Actors

**Trustor**

✅ Trustworthy

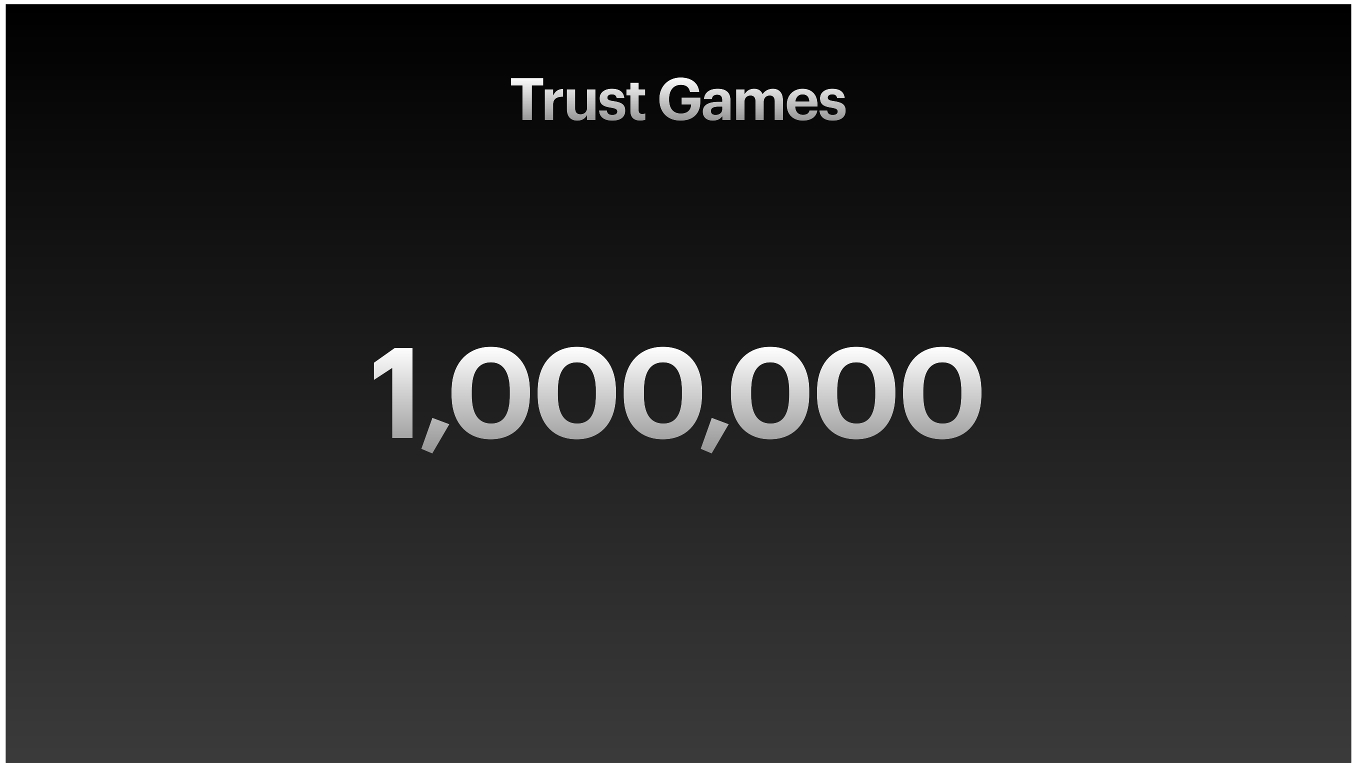
✅ Trusting

**Trustee**

❌ Trustworthy

❌ Trusting → ✅

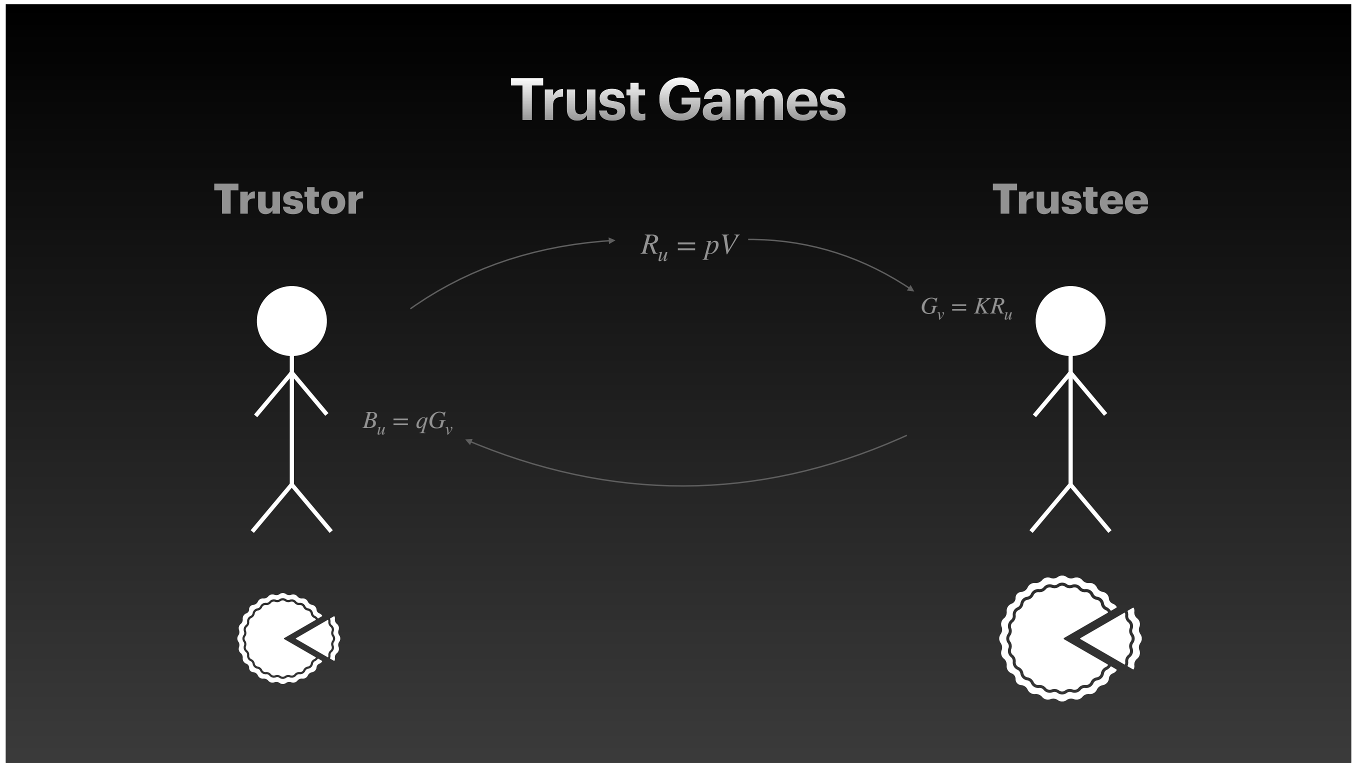
- The goal is to show them that they can trust you, thus make them trusting of you



## Trust Games

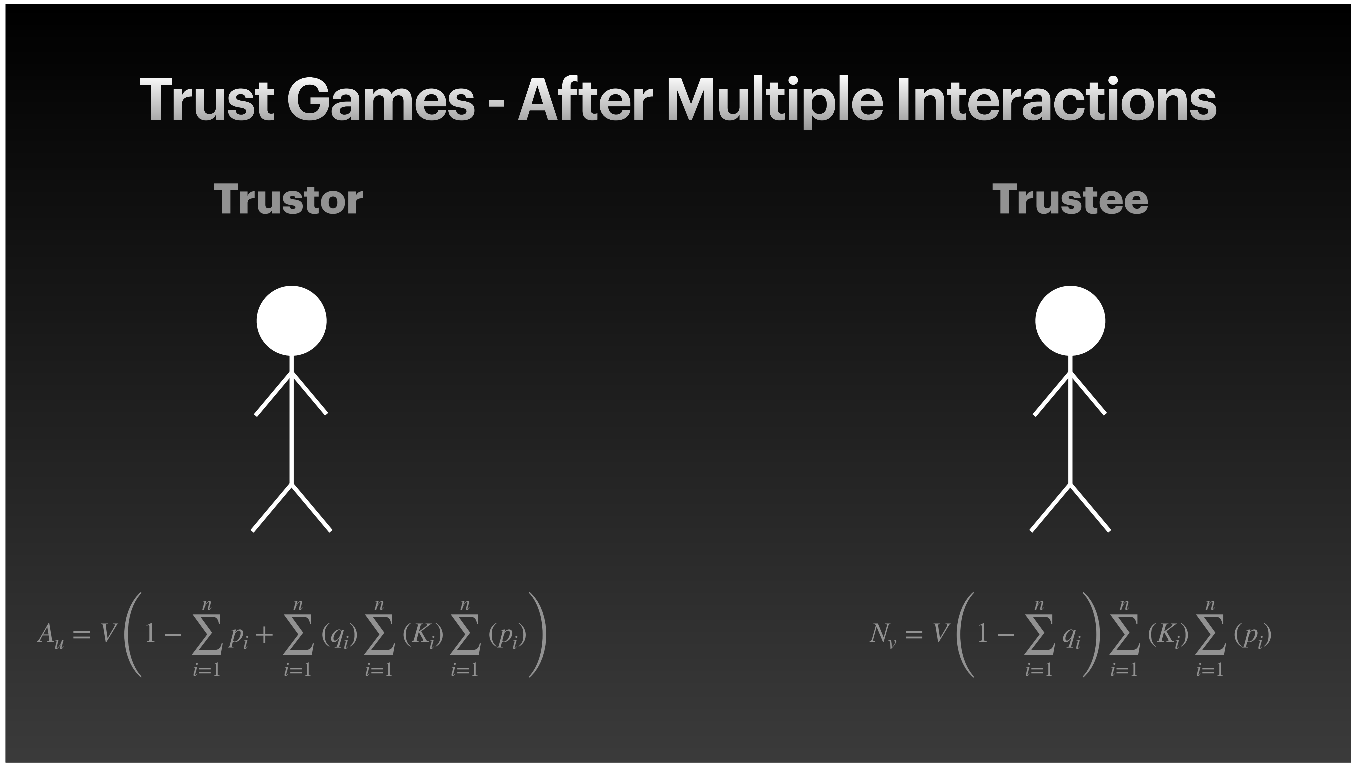
1,000,000

- Let us play a trust game. Like in the TV show “Who’s Line is It Anyway?”, the actual number of points don’t matter
- In our case, irrespective of the starting number of points, what we care about is how it fluctuates over time



# Trust Games

- Initial act of investing in building something of value. By selling, the trustor’s pie gets a little smaller
- *p* means partial remittance. *K* is the value magnification factor perceived by the trustee (e.g., a good book is something of value, worth more that the price you paid for it)
- When receiving the product/service, the trustee’s pie grow
- The portion *q* of the value sent back by the trustee reduces the size of its pie (less than it grew) and grows the trustor’s pie
- When the trustor interacts with multiple trustees, each trustee’s pies grow, and the accumulation of all the values sent back to the trustor, makes the investment worth their while



# Trust Games - After Multiple Interactions

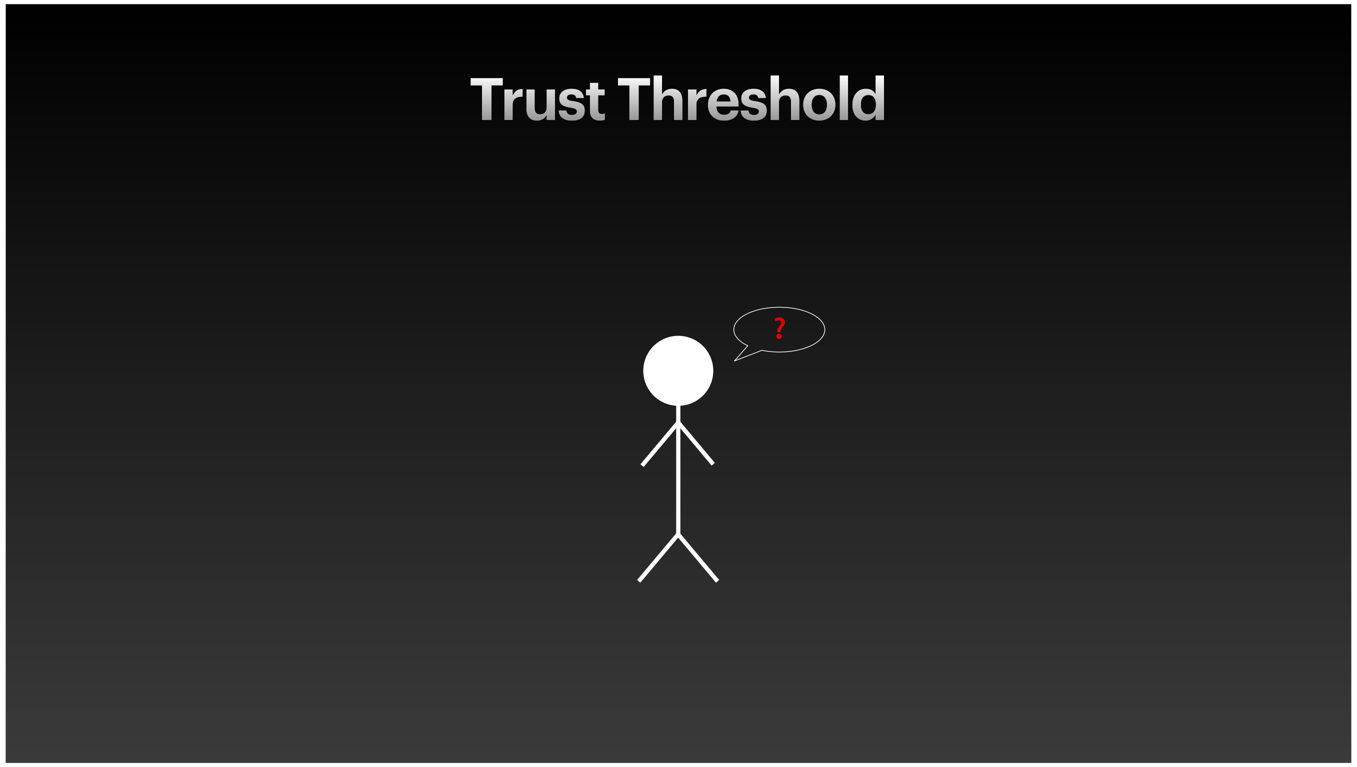
Trustor

Trustee

$$A_u = V\left(1 - \sum_{i=1}^{n} p_i + \sum_{i=1}^{n} (q_i) \sum_{i=1}^{n} (K_i) \sum_{i=1}^{n} (p_i)\right)$$

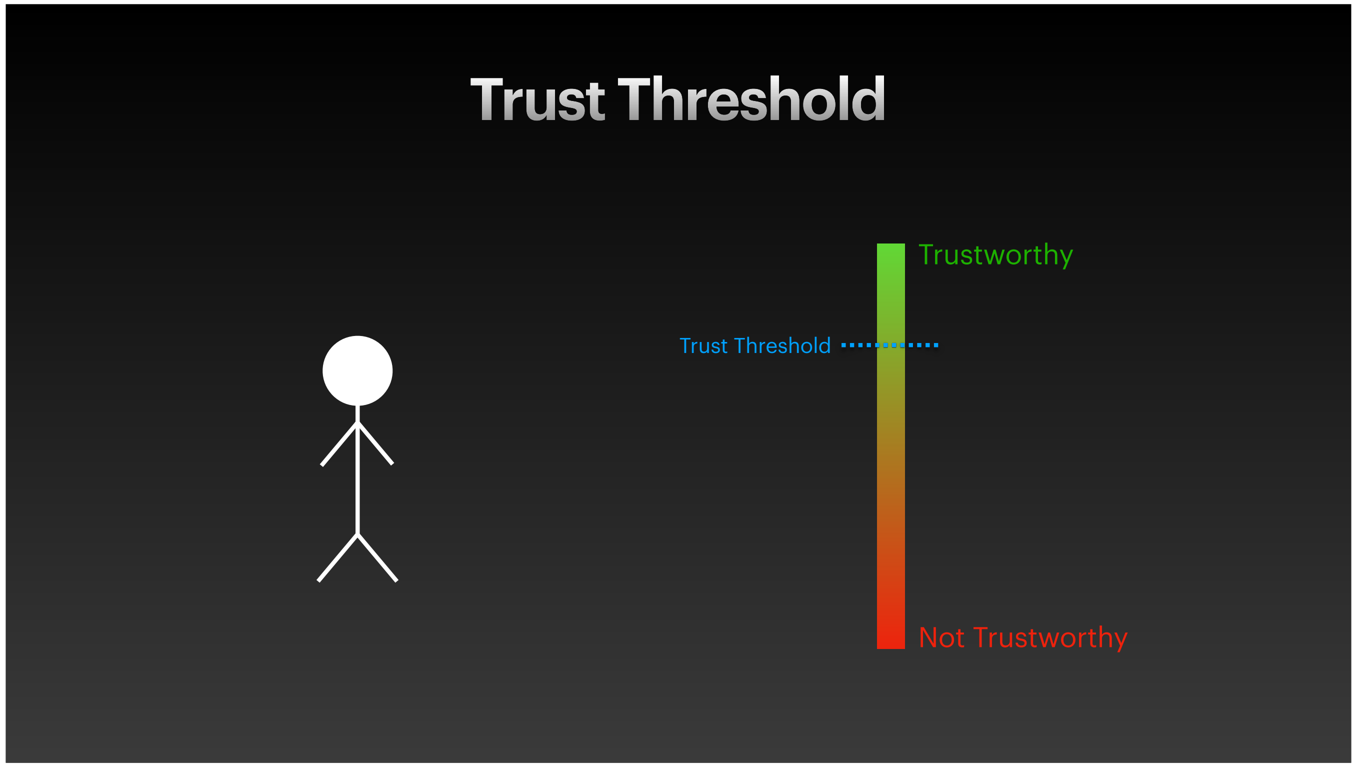
$$N_v = V\left(1 - \sum_{i=1}^{n} q_i\right) \sum_{i=1}^{n} (K_i) \sum_{i=1}^{n} (p_i)$$

- Expanding the Trust Games equations to multiple interactions
- There is no need to memorize the equations. But Remember Au (trustor’s accumulated value) and Nv (trustee’s net gains); we will refer to them later



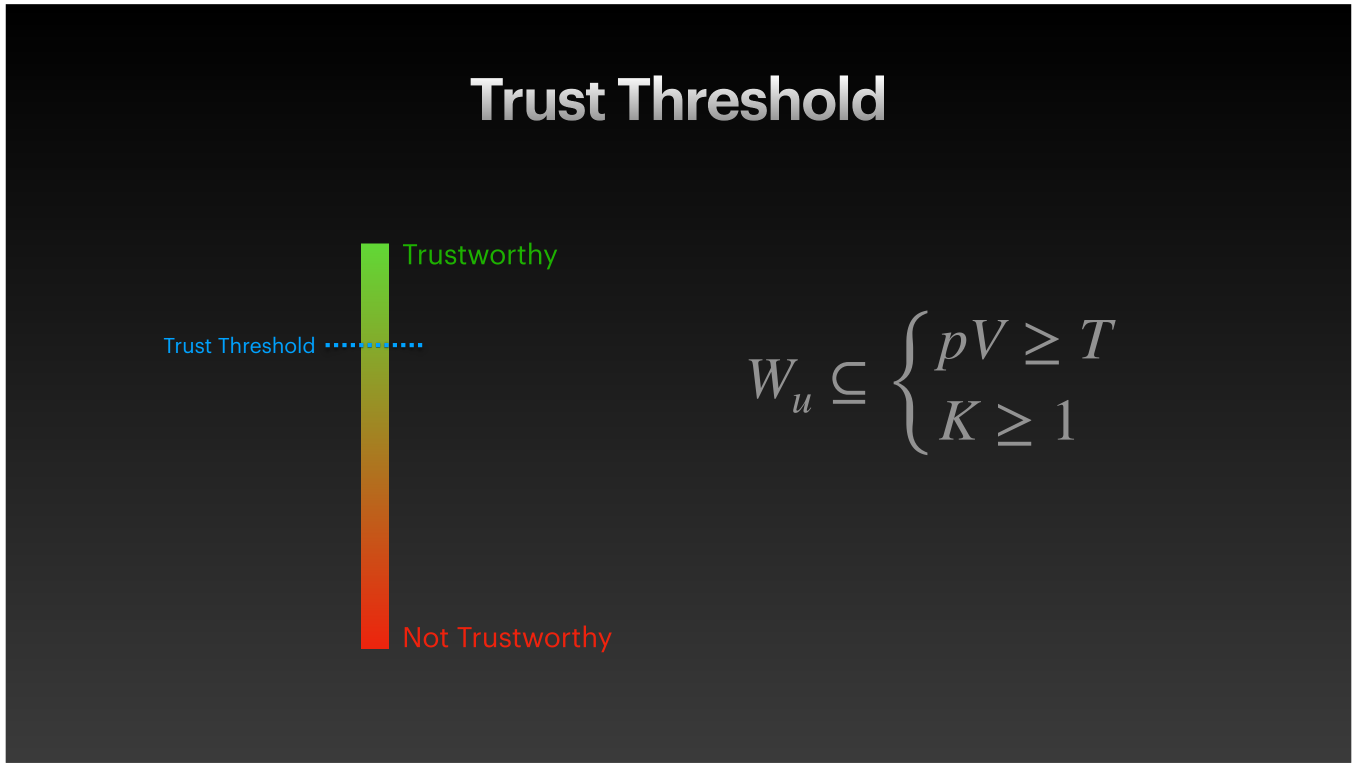
# Trust Threshold

- Ok, we’re quantifying trust, but the threshold to go from untrustworthy to trustworthy?



## Trust Threshold

- The threshold is established over time and determined on a case by case basis. Its numeric value is less important
- What is more important is that once trustworthiness reaches a certain level where Trust Games (trading) can take place, its value doesn't decay or fluctuates too widely over time



# Trust Threshold

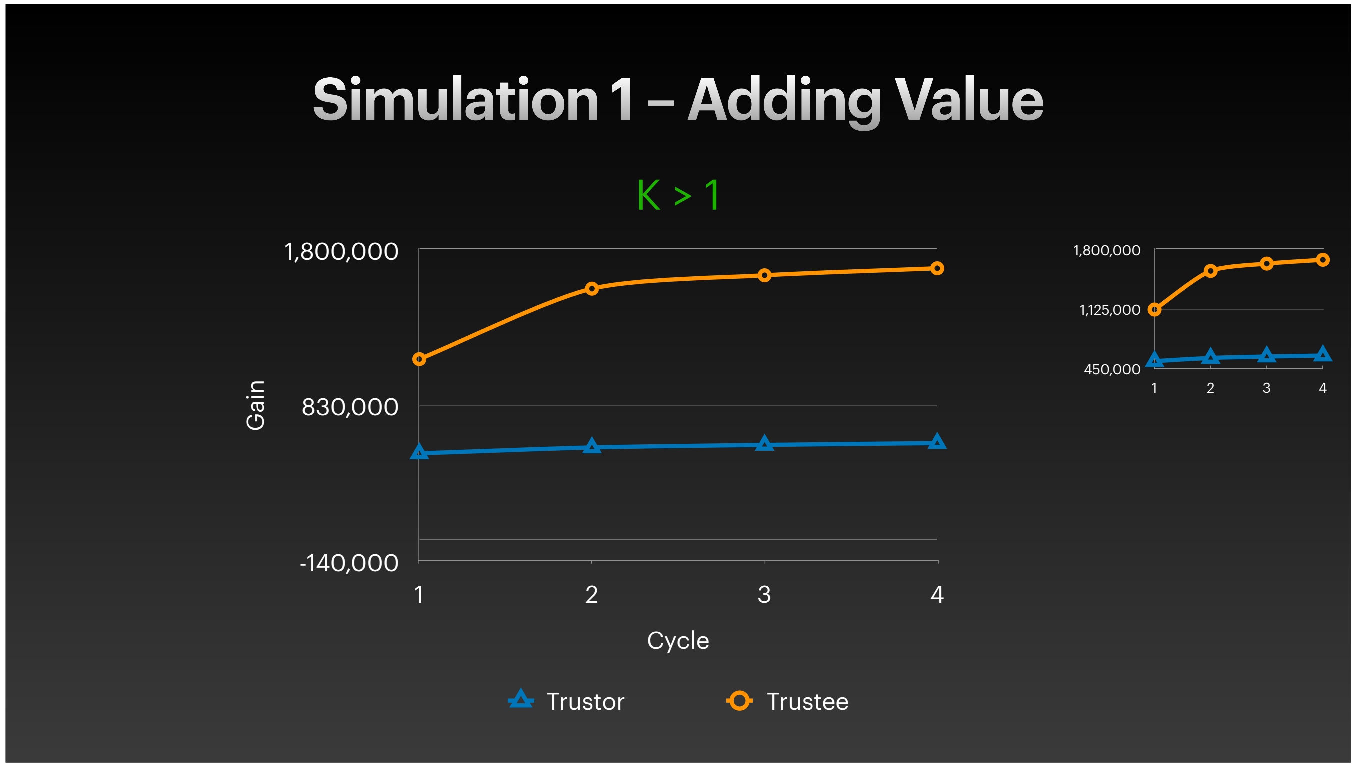
Trustworthy

Trust Threshold

Not Trustworthy

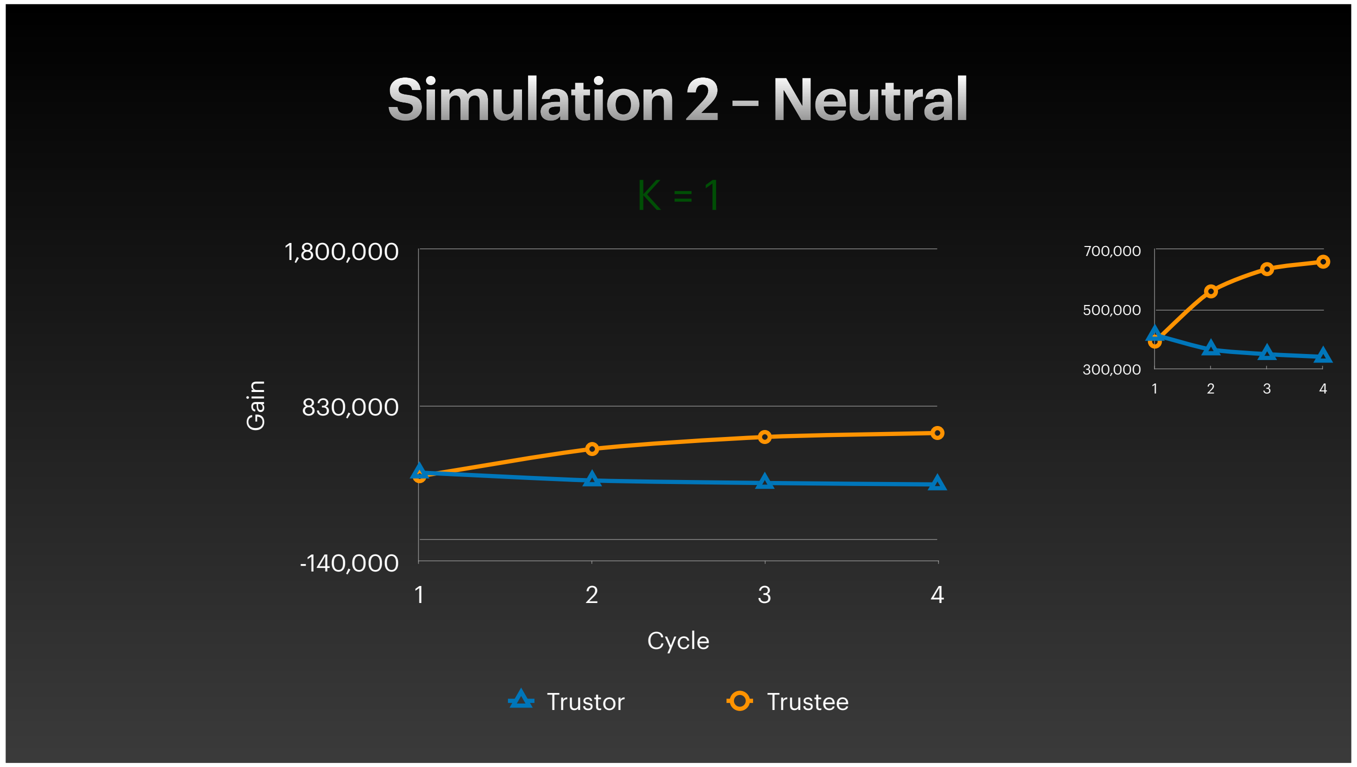
$$W_u \subseteq \begin{cases} pV \geq T \\ K \geq 1 \end{cases}$$

- Ideal circumstances for trust are when remittance is at or above the threshold and the magnification factor is greater than 1



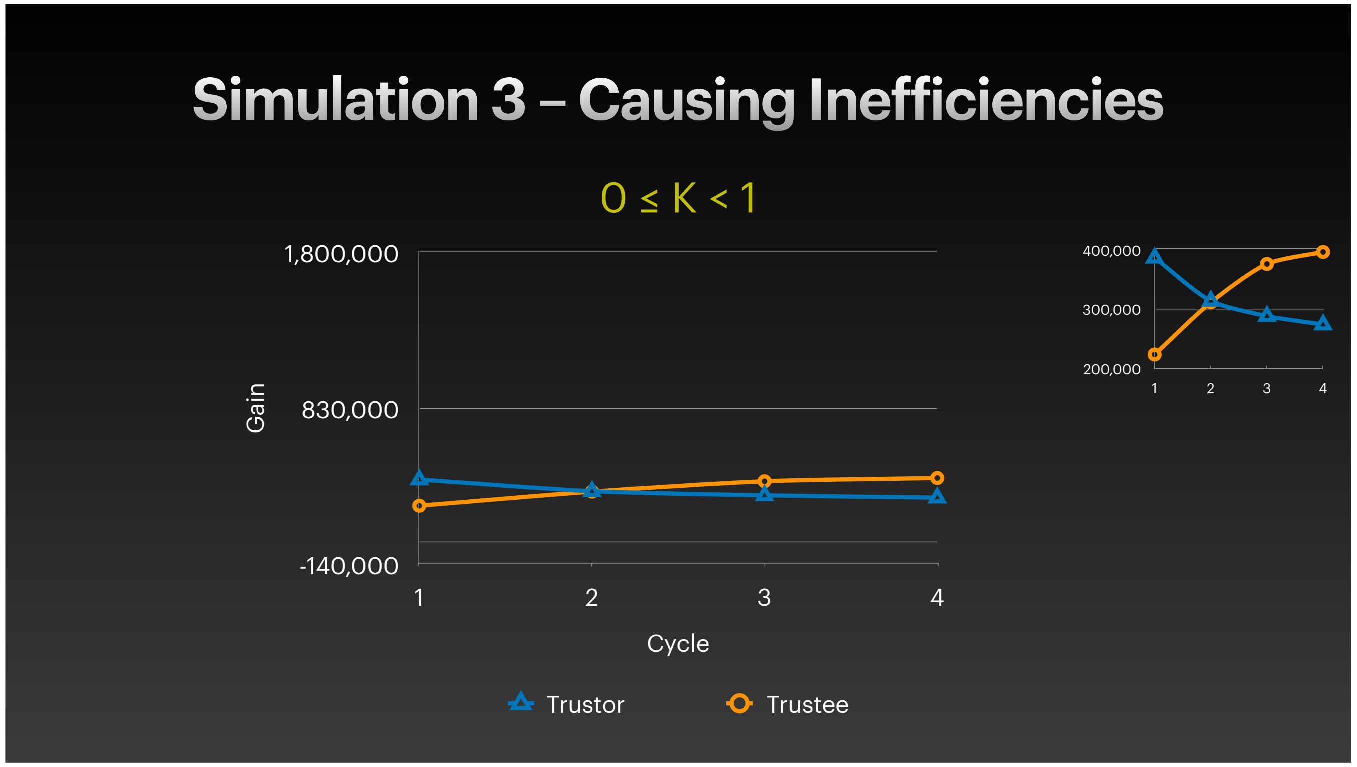
## Simulation 1 – Adding Value

- For this and the next simulation scenarios, the trustor begins with 1,000,000 points
- Trustor send remittances and trustee perceive the receive value with a magnification greater than 1
- Both trustor and trustee see their respective accumulated value and net gains grow up and to the right
- Notice the scale of the graph. On the top-right you can see a version of the graph plotted to the range of the simulation, rather than the global range



## Simulation 2 – Neutral

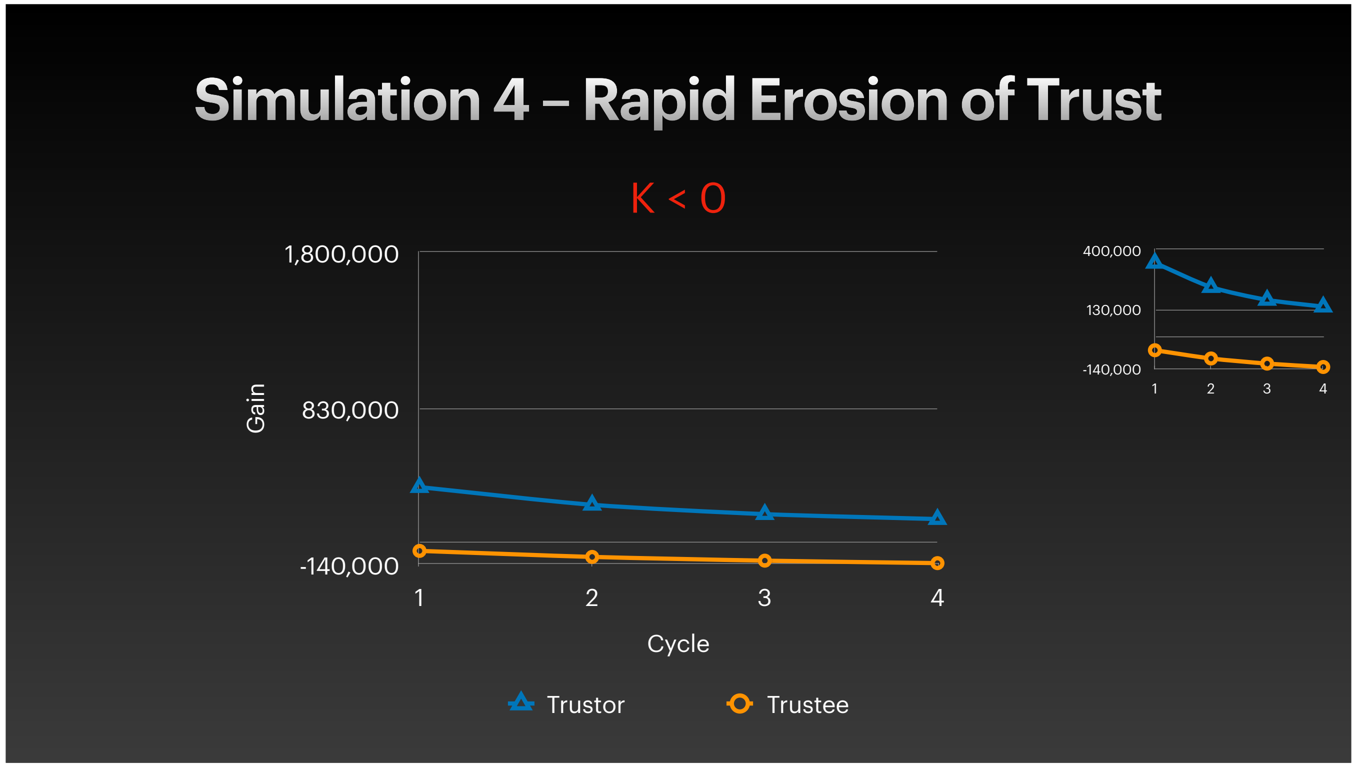
- Trustor send remittances and trustee perceive the receive value with a magnification equal to 1
- Trustee sees some modest net gains, but trustor experiences small depreciations (negative accumulations)



## Simulation 3 – Causing Inefficiencies

$0 \leq K < 1$

- Trustor send remittances and trustee perceive the receive value with a magnification greater than 0 and less than 1
- Trustee sees some modest net gains, but trustor experiences small depreciations (negative accumulations)
- This situation would be plausible and acceptable only during the development phase of a product, where a trustee would have accepted to be an early adopter of the service

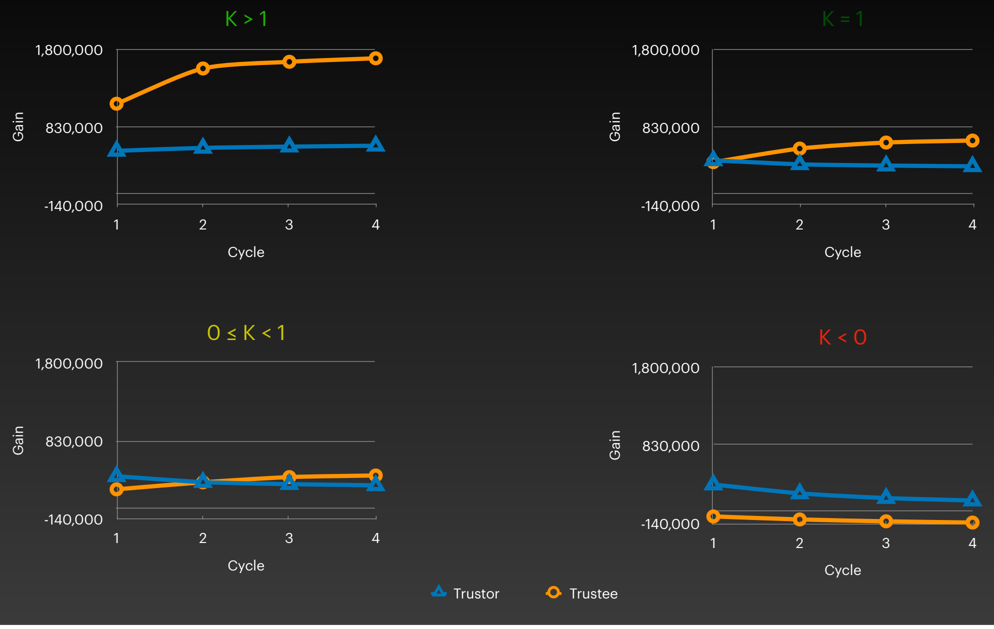


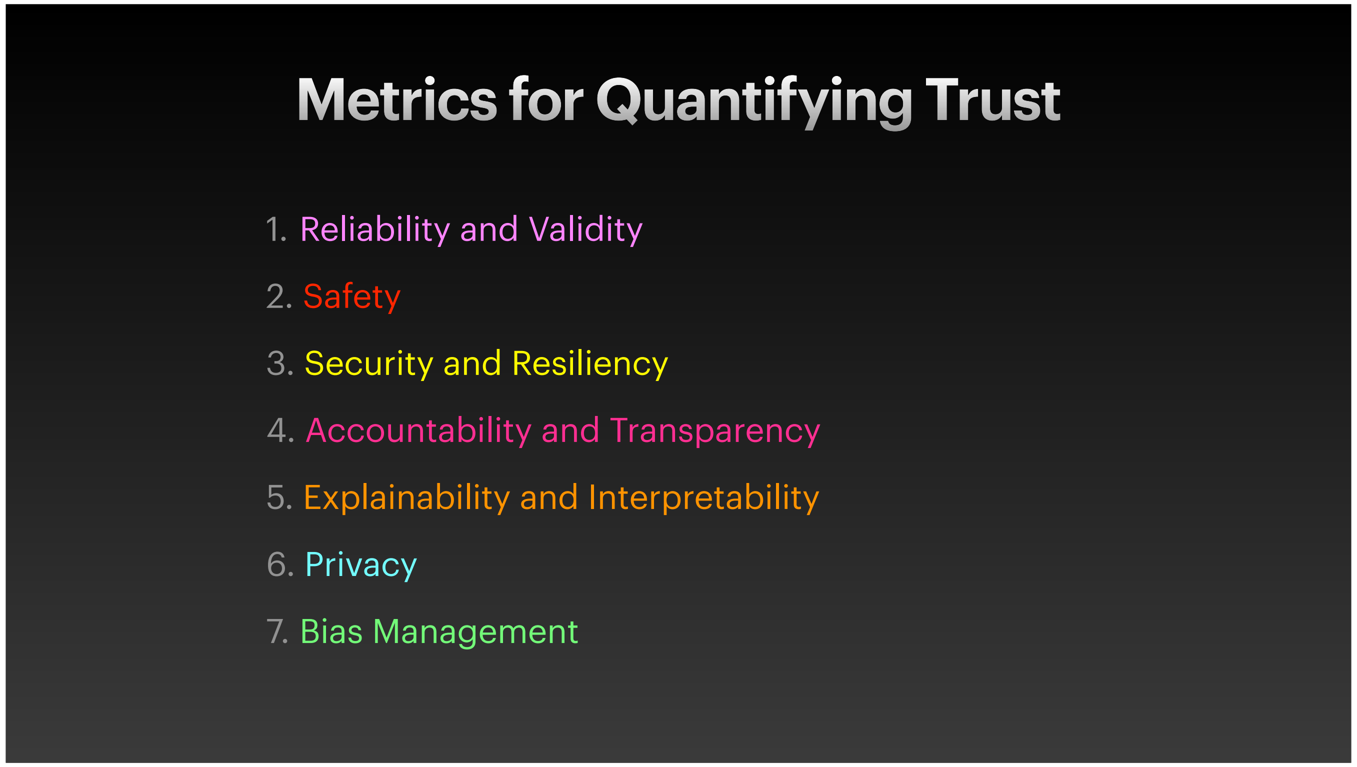
## Simulation 4 – Rapid Erosion of Trust

K < 0

- Trustor send remittances and trustee perceive the receive value with a magnification smaller than 0
- Both trustor and trustee see their respective accumulated value and net gains depreciate and decay
- Both are worse off with the product/service, compared to operating without it

# Simulations Side by Side

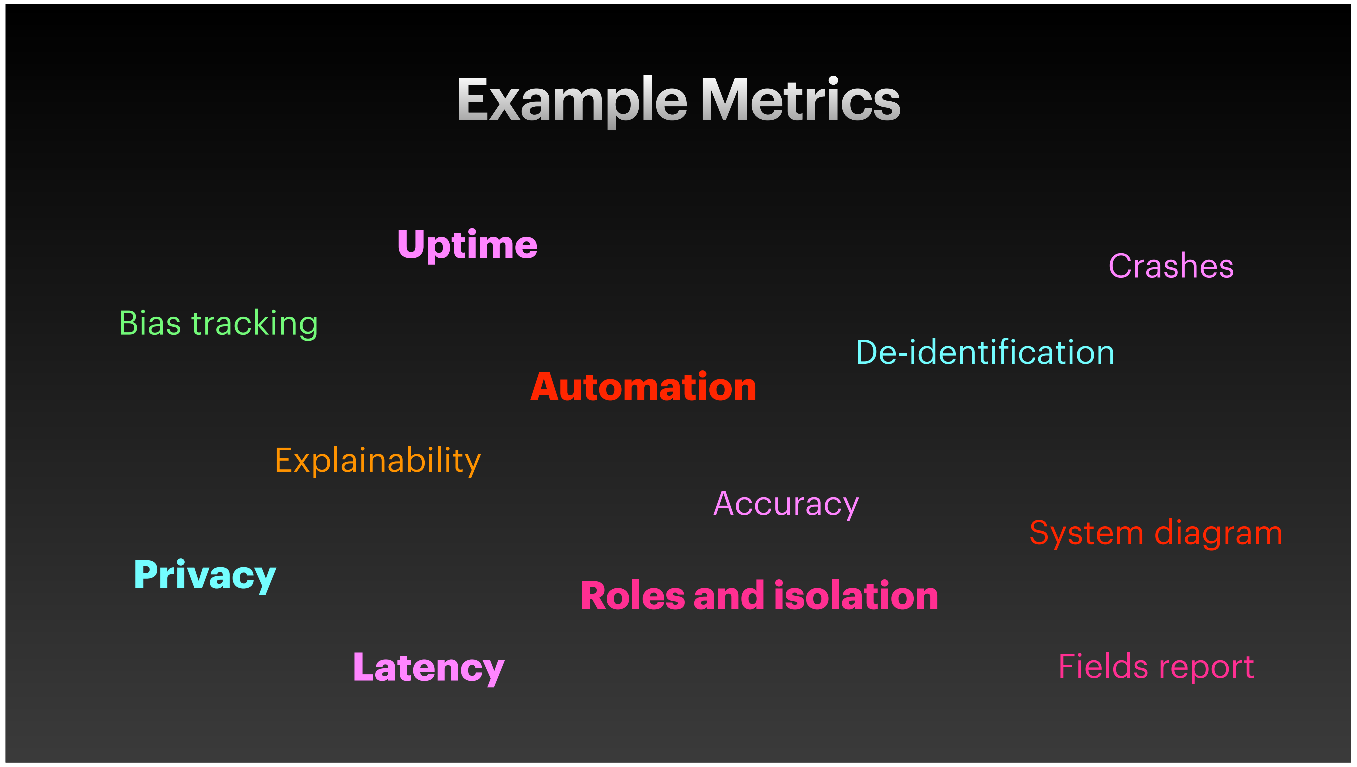




## Metrics for Quantifying Trust

1. Reliability and Validity
2. Safety
3. Security and Resiliency
4. Accountability and Transparency
5. Explainability and Interpretability
6. Privacy
7. Bias Management

- NIST, an American institute, published a study called “Artificial Intelligence Risk Management Framework (AI RMF).” There, they claim that there is a finite set of traits approximate to a good definition for a system to be trustworthy
- However, the categories are somewhat vague and qualitative only
- We think that it is possible to build upon this initial work and quantify each of the traits with measurable metrics



## Example Metrics

Uptime

Crashes

Bias tracking

De-identification

Automation

Explainability

Accuracy

System diagram

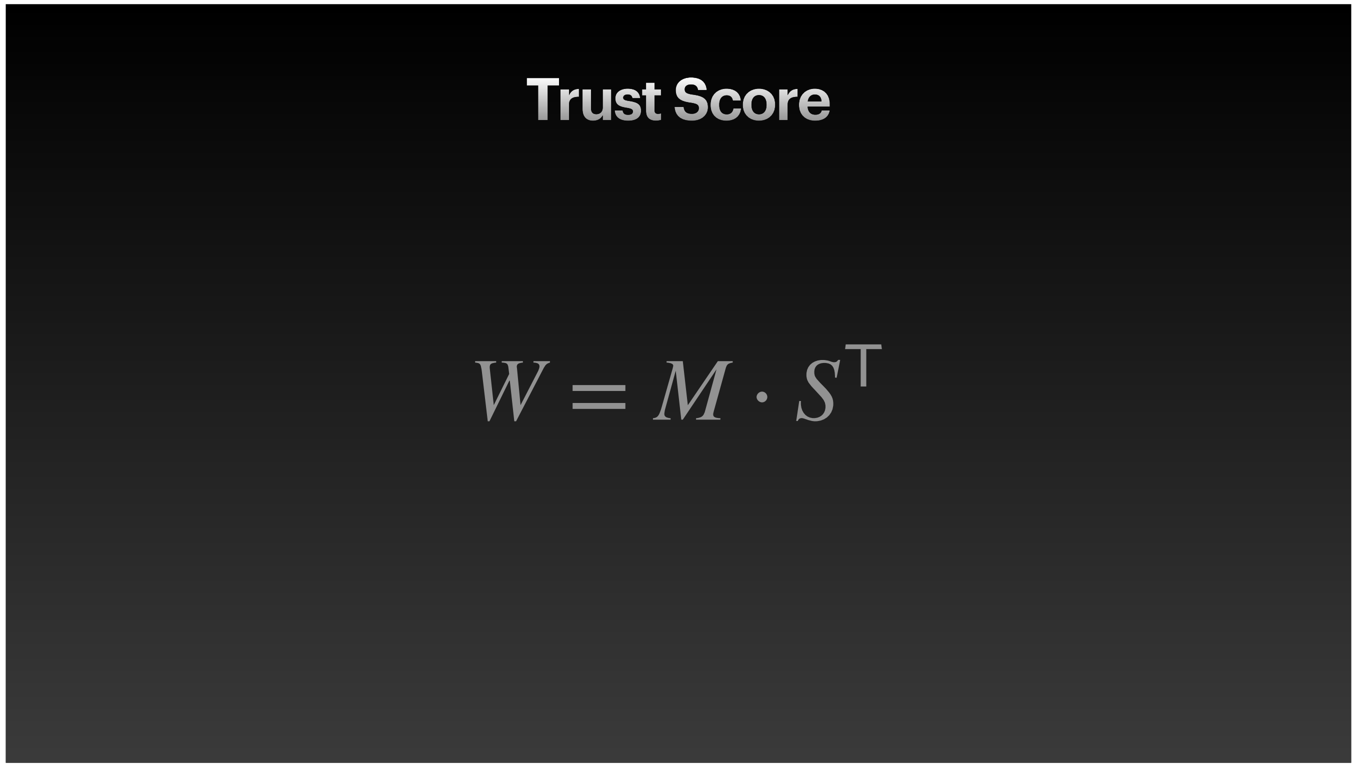
Privacy

Roles and isolation

Latency

Fields report

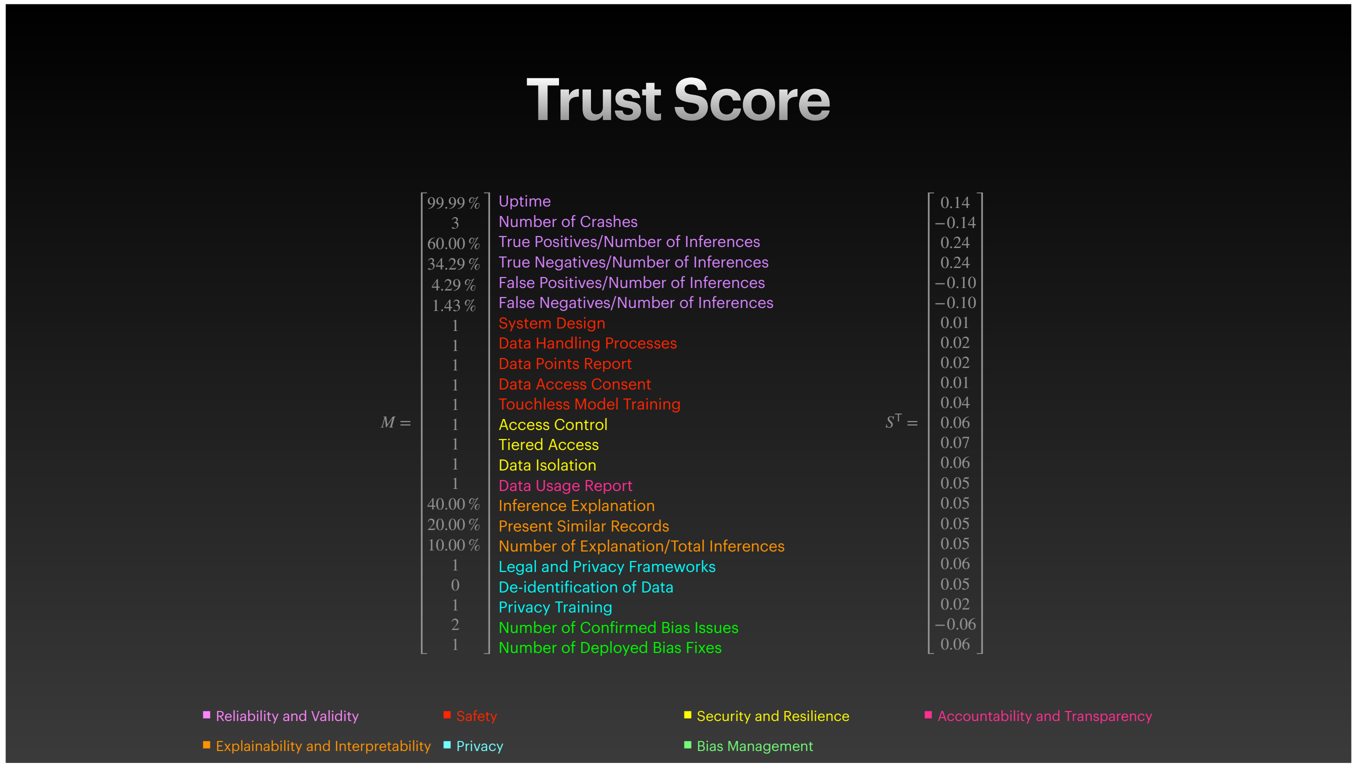
- The metrics are too many to fit on one slide. We selected a few to highlight over here
- The colors match the categories we just discussed



## Trust Score

$$W = M \cdot S^{\top}$$

- The idea behind the Trust Score is simple. A vector *M* containing each of the metrics, a stochastic vector *S* containing weights representing the importance, or contribution, of the retrospective metric
- The resulting dot product between the two vectors is the scalar *W*, which call Trust Score



# Trust Score

$$M = \left[\begin{array}{r} 99.99\,\% \\ 3 \\ 60.00\,\% \\ 34.29\,\% \\ 4.29\,\% \\ 1.43\,\% \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 40.00\,\% \\ 20.00\,\% \\ 10.00\,\% \\ 1 \\ 0 \\ 1 \\ 2 \\ 1 \end{array}\right] \begin{array}{l} \text{Uptime} \\ \text{Number of Crashes} \\ \text{True Positives/Number of Inferences} \\ \text{True Negatives/Number of Inferences} \\ \text{False Positives/Number of Inferences} \\ \text{False Negatives/Number of Inferences} \\ \text{System Design} \\ \text{Data Handling Processes} \\ \text{Data Points Report} \\ \text{Data Access Consent} \\ \text{Touchless Model Training} \\ \text{Access Control} \\ \text{Tiered Access} \\ \text{Data Isolation} \\ \text{Data Usage Report} \\ \text{Inference Explanation} \\ \text{Present Similar Records} \\ \text{Number of Explanation/Total Inferences} \\ \text{Legal and Privacy Frameworks} \\ \text{De-identification of Data} \\ \text{Privacy Training} \\ \text{Number of Confirmed Bias Issues} \\ \text{Number of Deployed Bias Fixes} \end{array} \qquad S^{\mathsf{T}} = \left[\begin{array}{r} 0.14 \\ -0.14 \\ 0.24 \\ 0.24 \\ -0.10 \\ -0.10 \\ 0.01 \\ 0.02 \\ 0.02 \\ 0.01 \\ 0.04 \\ 0.06 \\ 0.07 \\ 0.06 \\ 0.05 \\ 0.05 \\ 0.05 \\ 0.05 \\ 0.06 \\ 0.05 \\ 0.02 \\ -0.06 \\ 0.06 \end{array}\right]$$

■ Reliability and Validity ■ Safety ■ Security and Resilience ■ Accountability and Transparency

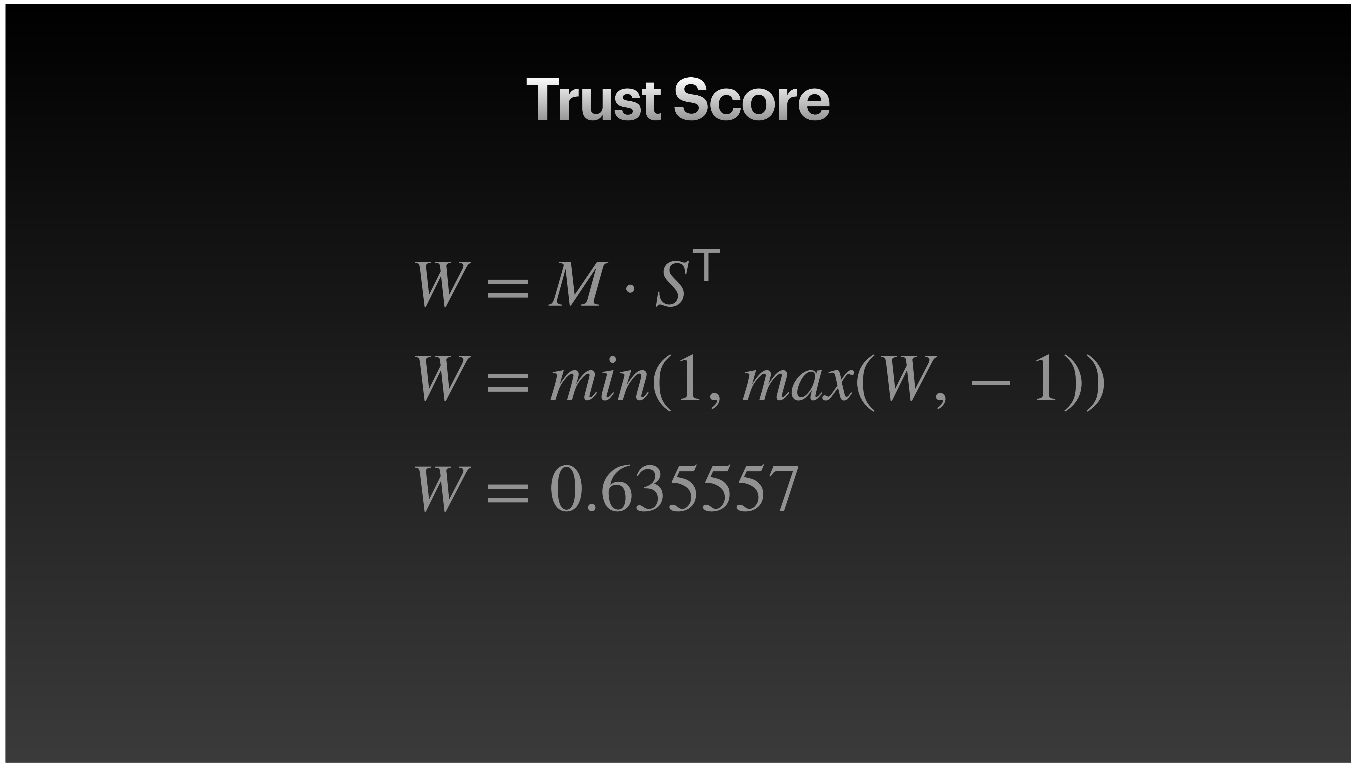
■ Explainability and Interpretability ■ Privacy ■ Bias Management

- This is an example showing the all the metrics and suggested weights

# Trust Score

$$M = \begin{bmatrix} 99.99\,\% \\ 3 \\ 60.00\,\% \\ 34.29\,\% \\ 4.29\,\% \\ 1.43\,\% \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 40.00\,\% \\ 20.00\,\% \\ 10.00\,\% \\ 1 \\ 0 \\ 1 \\ 2 \\ 1 \end{bmatrix}$$

$$S = [0.14 \quad -0.14 \quad 0.24 \quad 0.24 \quad -0.10 \quad -0.10 \quad 0.01 \quad 0.02 \quad 0.02 \quad 0.01 \quad 0.04 \quad 0.06 \quad 0.07 \quad 0.06 \quad 0.05 \quad 0.05 \quad 0.05 \quad 0.05 \quad 0.06 \quad 0.05 \quad 0.02 \quad -0.06 \quad 0.06]$$



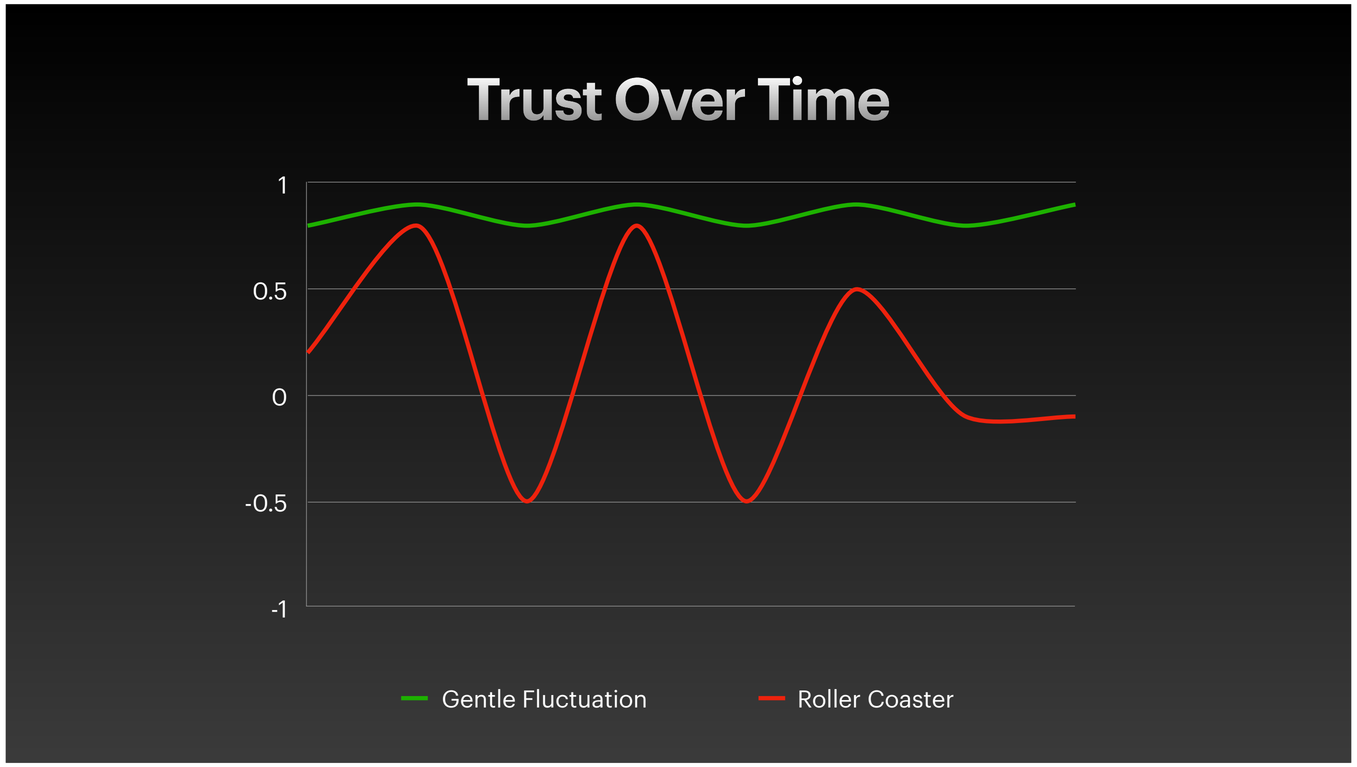
## Trust Score

$$W = M \cdot S^{\top}$$

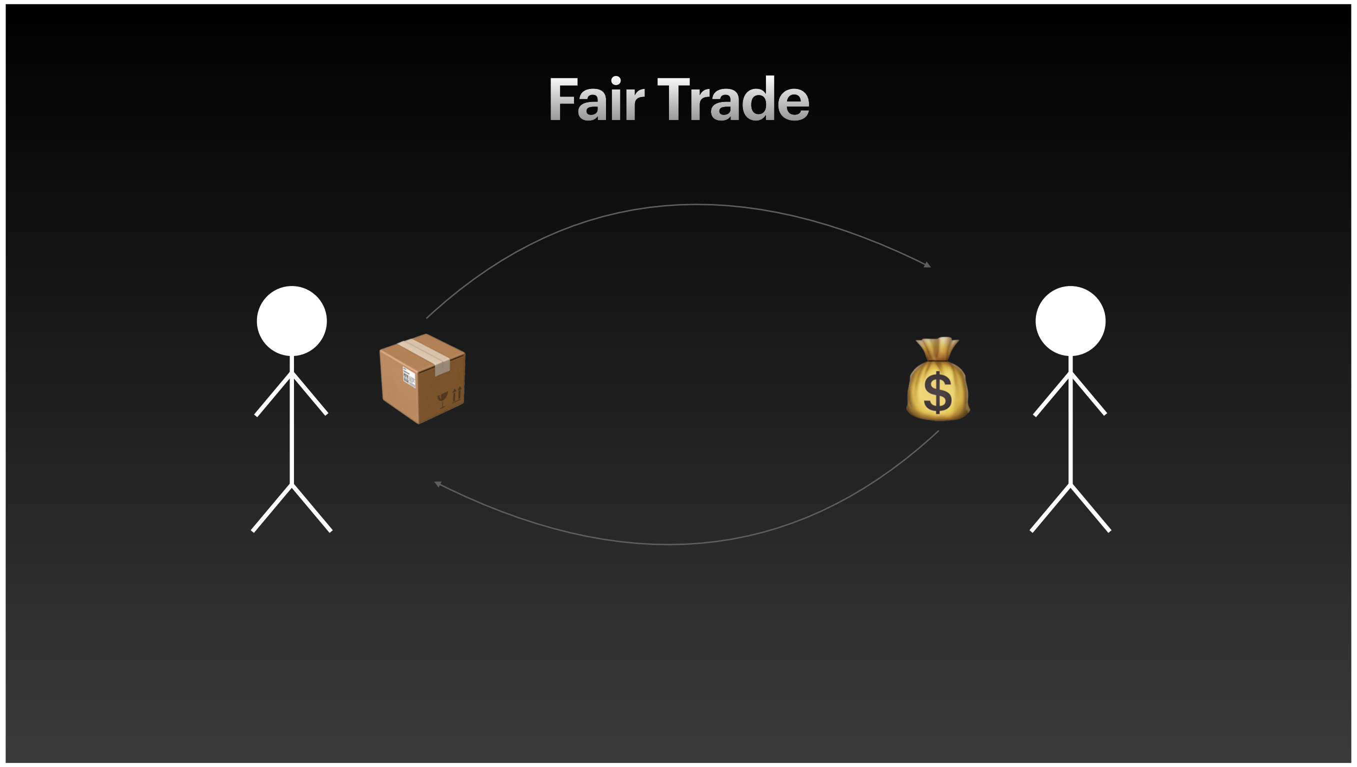
$$W = min(1, max(W, -1))$$

$$W = 0.635557$$

- In addition to computing the trust score, we want to limit it to the range between -1 and 1 (for practical purposes and better understanding from the public in general)
- -1 is the worst possible score, and 1 is the optimal score
- For the example provided here, the Trust Score is 0.635

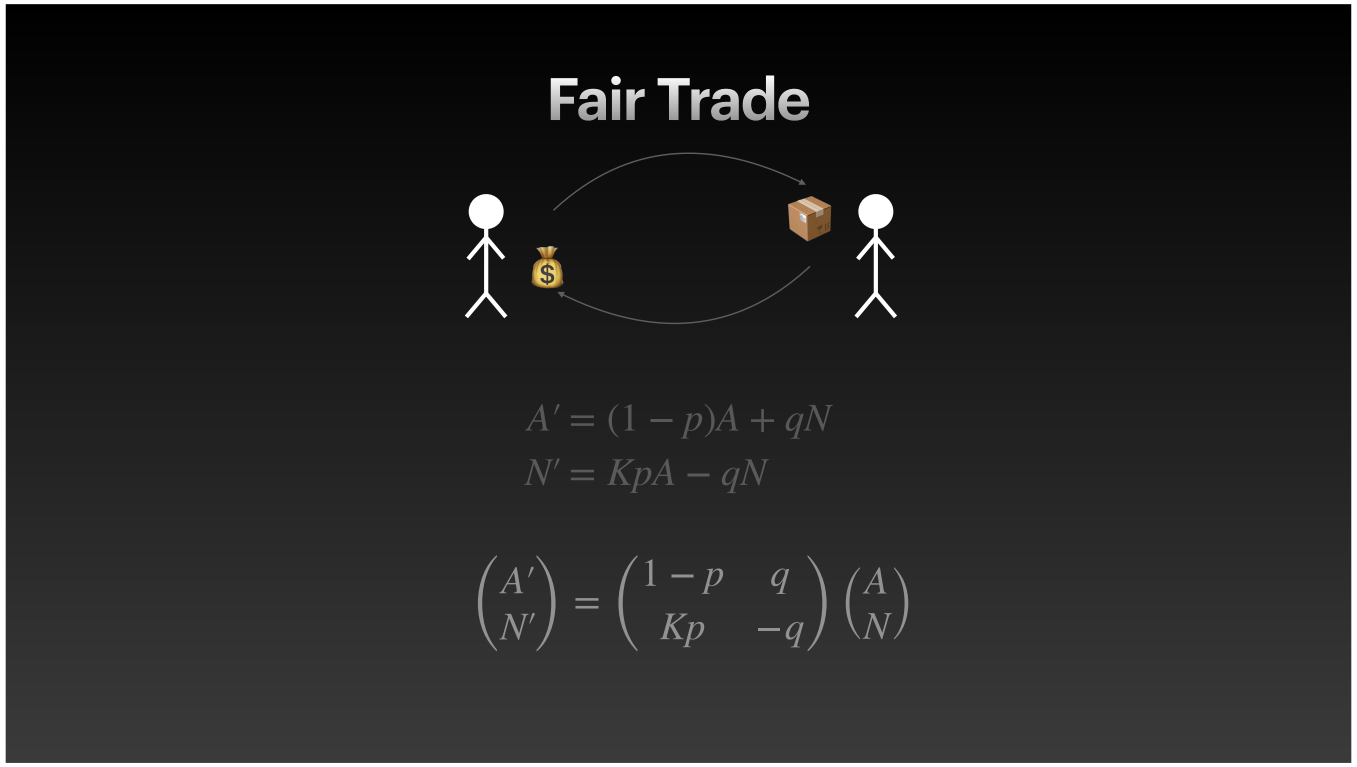


- It is expected that trust will fluctuate over time, but it should be gentle
- Large fluctuations like a roller coaster will erode trust and possibly lead to a break up



## Fair Trade

- We discussed trust, quantifying it, establishing a threshold, and multiple interactions. But how do we know if the trade between trustor and trustee is fair?

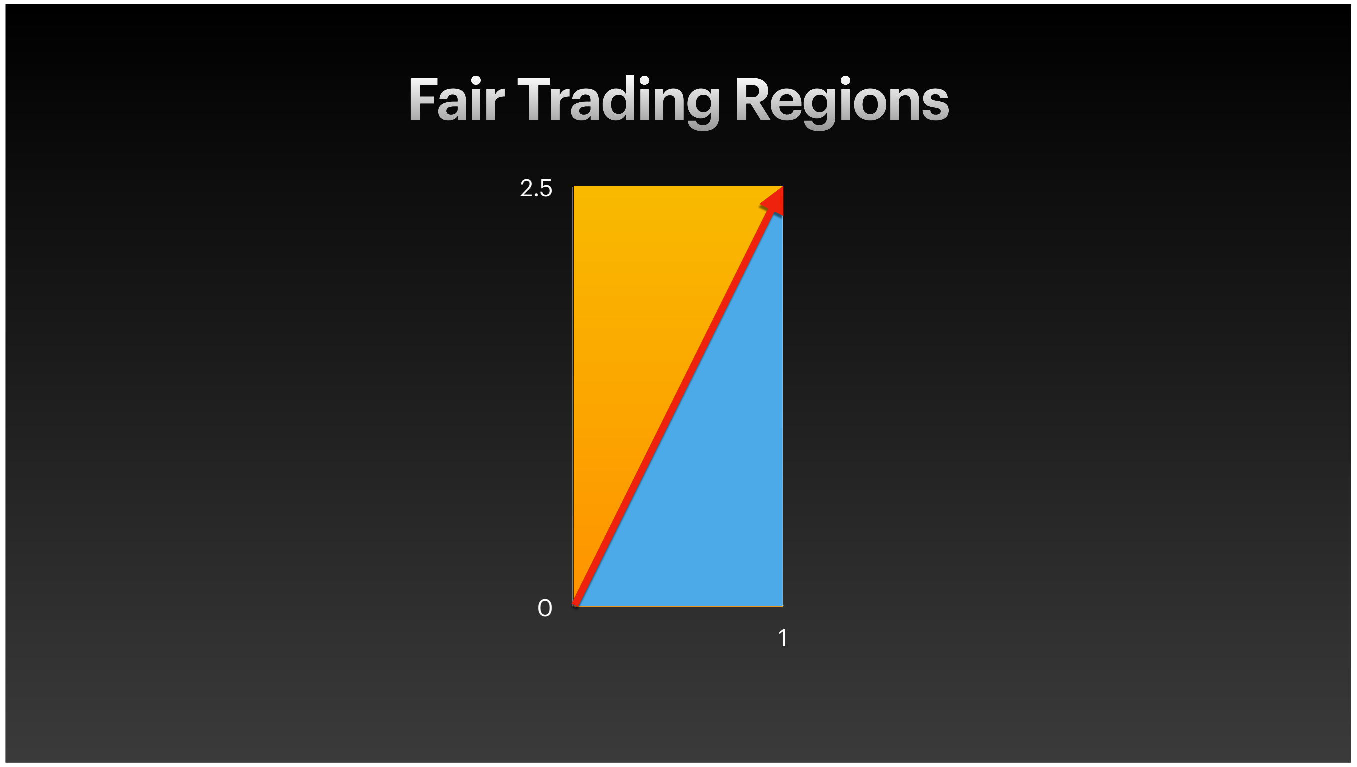


## Fair Trade

$$A' = (1 - p)A + qN$$
$$N' = KpA - qN$$

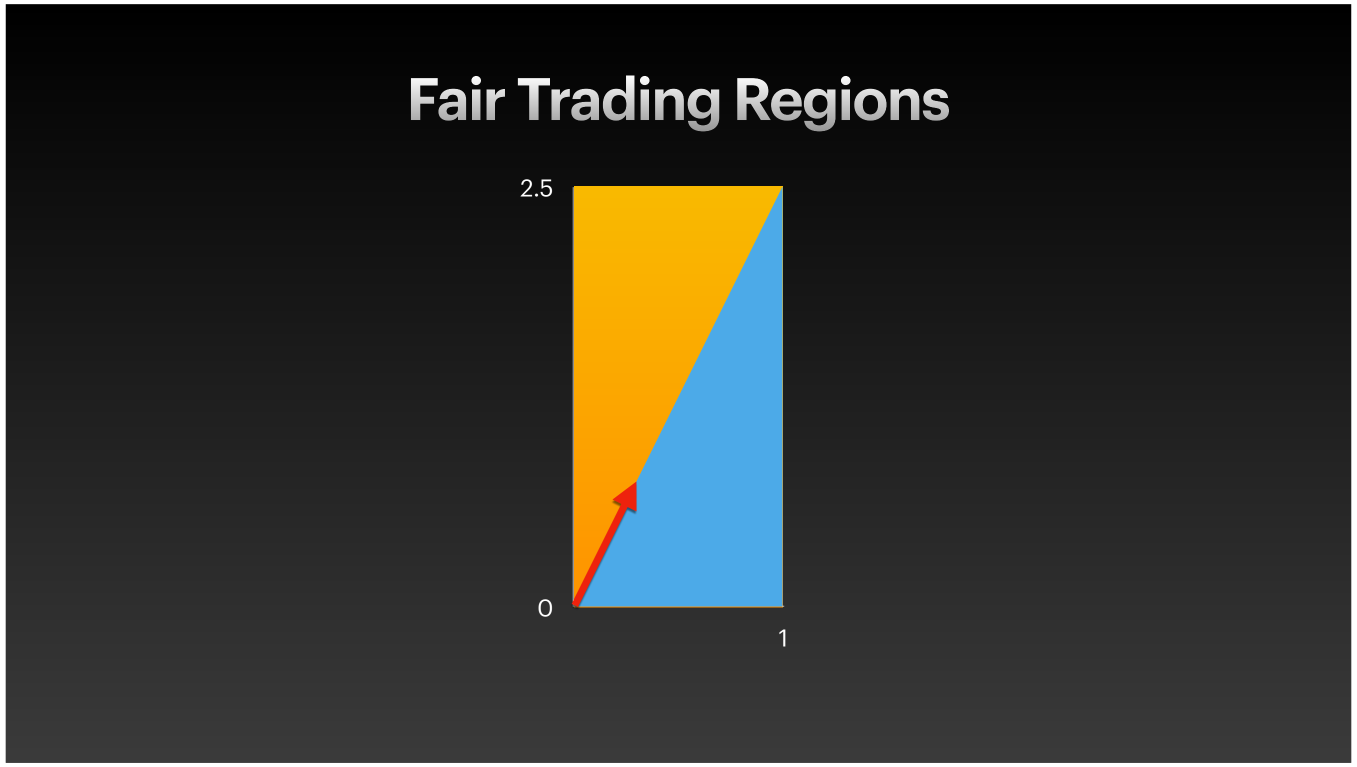
$$\begin{pmatrix} A' \\ N' \end{pmatrix} = \begin{pmatrix} 1 - p & q \\ Kp & -q \end{pmatrix} \begin{pmatrix} A \\ N \end{pmatrix}$$

- After multiple interactions the trustor will have accumulated a value *A*, and trustee will have had a net gain *N*
- The results of the next interaction can be predicted using the following equations, where *A’* and *N’* are the next state upon completion of the interaction
- And by expressing the system of equations as a matrix, we are able to compute its eigenvector



# Fair Trading Regions

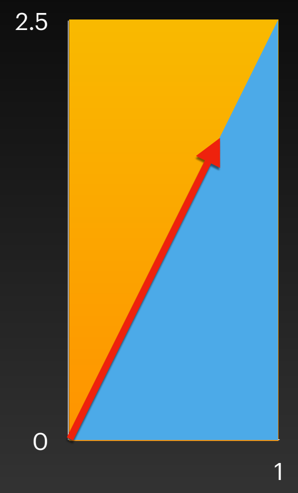
- An eigenvector will only scale, no matter what linear transformation is applied to it
- The area highlighted in orange represents the value for the trustor, and the one in blue represents the value for the trustee



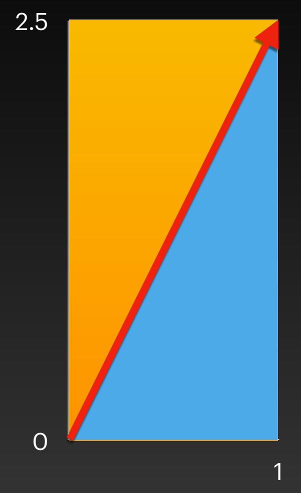
# Fair Trading Regions

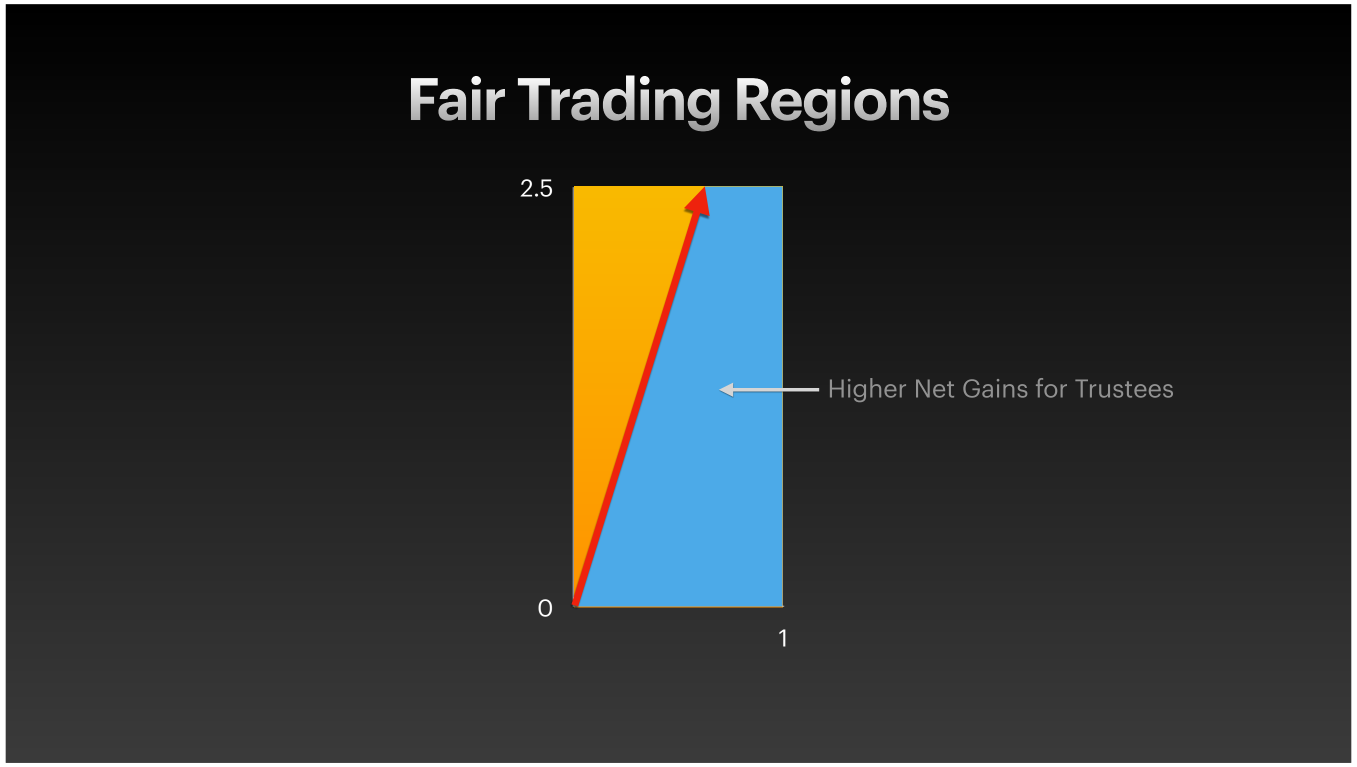
- When fair trading is taking place, the accumulated value for the trustor and the net gains for trustees expand and contract proportionally

# Fair Trading Regions



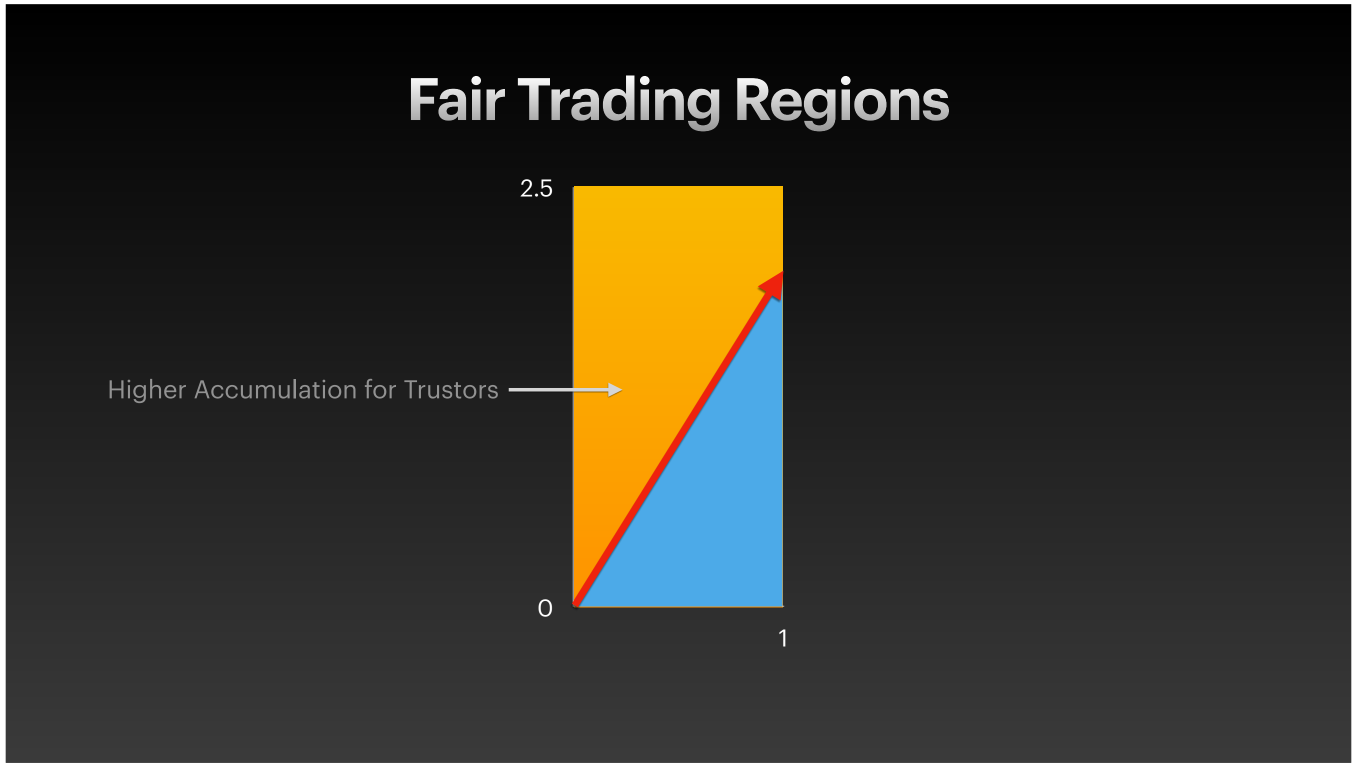
# Fair Trading Regions





## Fair Trading Regions

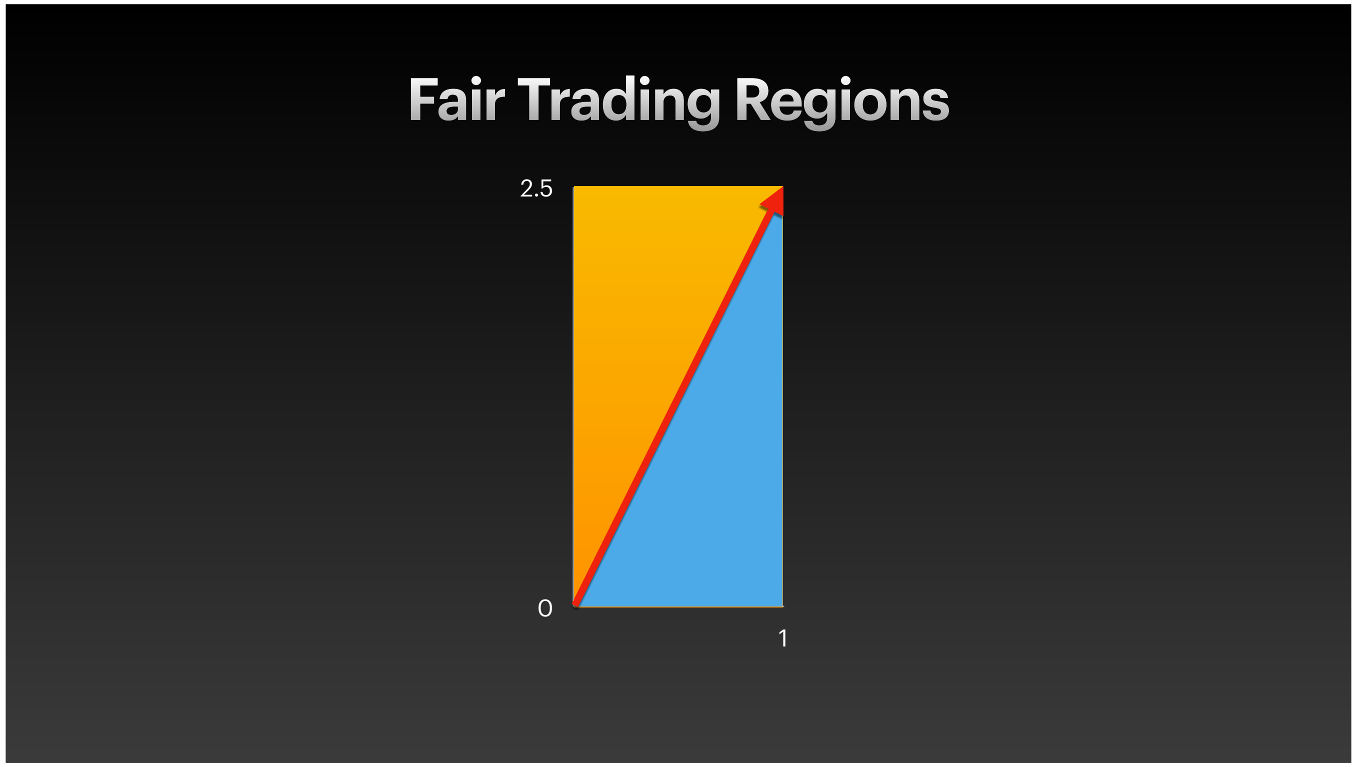
- If a trustee is demanding more and more and paying less and less, the system becomes unbalanced to its favor, at the expense of the trustor



## Fair Trading Regions

Higher Accumulation for Trustors

- On the other hand, if a trustor is charging more and more and delivering less and less, the system becomes unbalanced to its favor, at the expense of trustees



# Fair Trading Regions

- Trust can be built and is quantifiable
- Trading can be fair

# Let’s Talk. Here and Online.

Dalmo Cirne
https://dalmocirne.com

Veena Calambur
https://www.linkedin.com/in/veena-calambur/

IARIA 2024