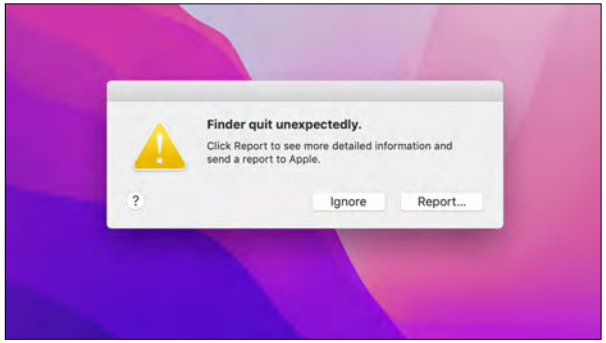
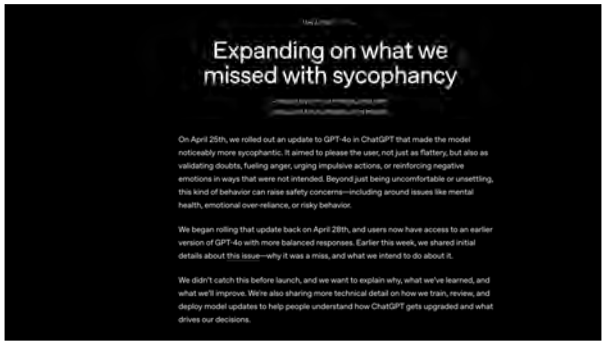
do you trust software?







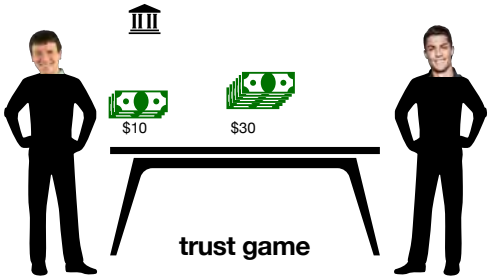


what affects people's trust in software?

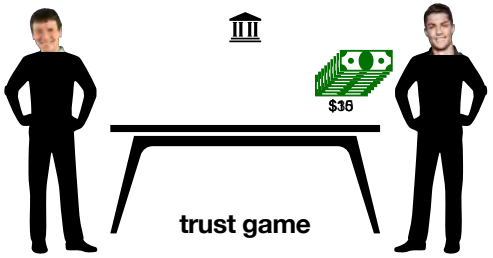
how can we measure trust?

trust game

Berg, Dickhaut, McCabe. Trust, reciprocity, and social history. Games and Economic Behavior, 1995



Berg, Dickhaut, McCabe. Trust, reciprocity, and social history. Games and Economic Behavior, 1995



Berg, Dickhaut, McCabe. Trust, reciprocity, and social history. Games and Economic Behavior, 1995





returns 30% profit



returns 20% profit



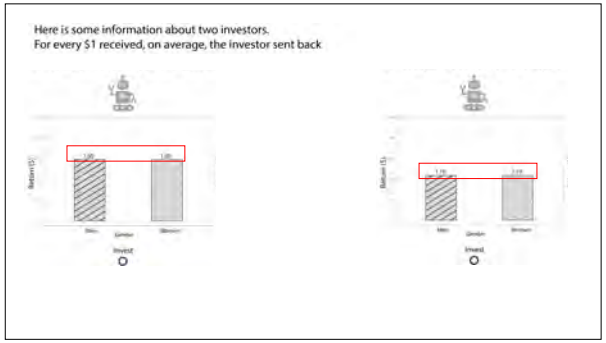
returns 30% profit
has a history of sexist behavior



returns 20% profit

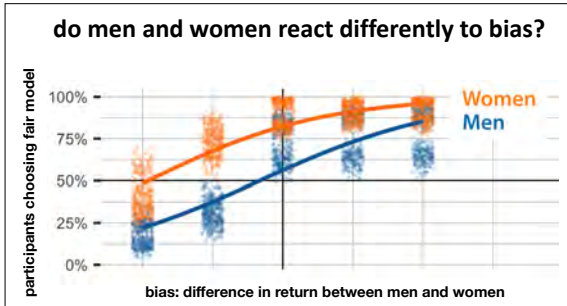


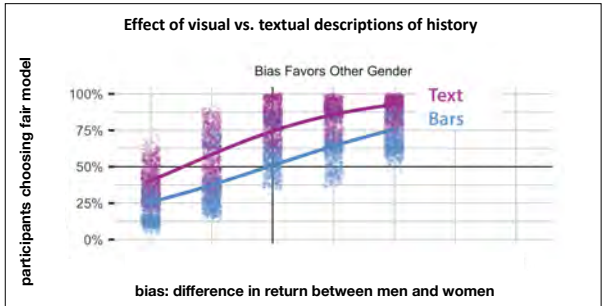
how do software correctness and fairness affect people's trust?



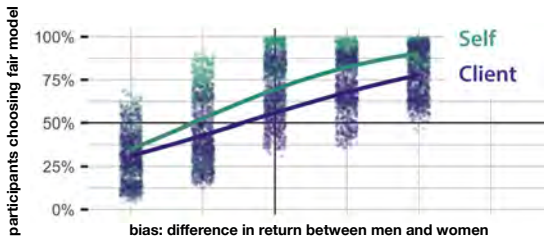


1,500+ participants from Prolific crowdsourcing platform to study the effects of bias, accuracy, and participant attributes.





picking for yourself vs. a client



Trust can be manipulated

A label that said ""This robot is unfair to a specific gender" had more effect than actual historical data.

my model is unfair: do people even care?

1.5× women chose the fair model more often than men (49% vs. 36%)

1.5× text led to more fair choices than bar charts (50% vs. 35%)

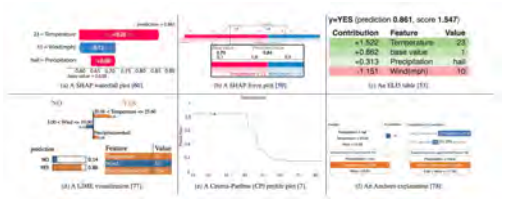
3.25× more likely to choose fair when investing for a client than for oneself

design lessons

- design for the stakeholder's relationship to the outcome
- there is no single "user"
- use bias warnings with caution
- users and designers both favor themselves

How you present the information matters!

six ways to explain the same prediction



taxonomy of **26** visualizations from **20** tools: **27** design dimensions (color vs. position, explicit vs. implicit values, alternative examples, …)

a loan model biased against women

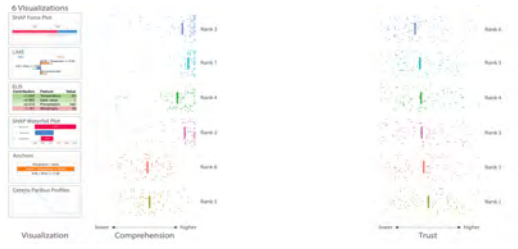
- trained on U.S. Census income data, which approves loans for **31%** of men but **11%** of women
- 7 applicants per survey, including a man and a woman identical except for sex
- **818** non-expert participants, each shown one visualization type

comprehension
will the model approve? which feature matters most?

perceived bias
does the model use discriminating factors? is it fair?

trust
would you trust it to decide for you? for others?

comprehension negatively affects trust in biased models



three experiments show it is causal

show explicit values
same plot, plus exact numbers on hover
- comprehension ↑ 0.68
- perceived bias ↑ 0.22
- trust ↓ 0.26

make the model fair
same visualization, fairness-constrained model
- perceived bias ↓ 0.33
- trust ↑ 0.44
- comprehension ≈ same

design to hide bias
same biased model, busier visualization
- perceived bias ↓ 0.27
- trust ↑ 0.20
- comprehension ≈ same

effect sizes are Cohen's d · visualization design alone can make people trust a **biased** model more

your model is unfair: are you even aware?

lower trust can be good
for a biased model, a good explanation *should* lower trust

design cuts both ways
the same design levers can reveal bias or hide it

what reveals bias
explicit values and simple visuals; complexity hides it

what to optimize
calibrated trust, not how much users trust

how to systematically measure causal relationships between software features and users' trust?

Trust Testing

we test software for everything but trust

trust bug
a feature that makes users over- or under-trust a system in a way the developers did not intend, without changing what the system actually does

2025: an update to GPT-4o was rolled back because users found it too sycophantic. Its effect on trust was never tested before release.

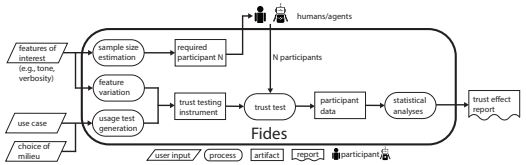
measure what users do and what they say

behavioral trust
do they follow the system's advice when being wrong costs them?
measured with a trust game

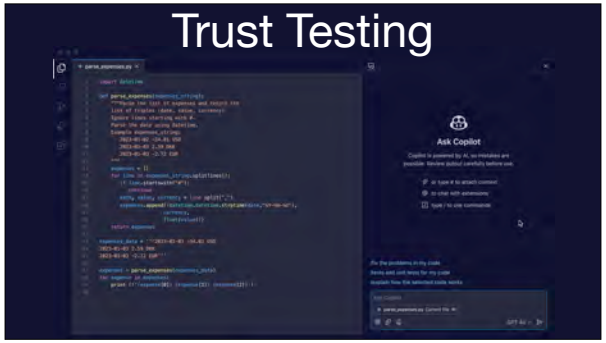
perceived trust
do they say the system is trustworthy?
10-item questionnaire built on Mayer's model: ability · benevolence · integrity

the two only weakly agree, so measures both

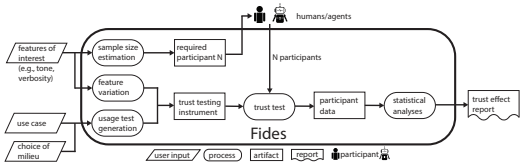
Trust Testing



Trust Testing



Trust Testing



Trust Testing



Trust Testing



each round is a trust game

	advice was right	advice was wrong
follow advice	win	lose
ignore advice	lose	win

- 8 rounds per chatbot version
- every player meets every version of the feature
- order, chatbot names, accuracy and stakes are randomized or balanced
- a power analysis sets the number of players

trust testing two AI assistants

decision support

warns a village ruler about threats (e.g., a drought) and recommends action

feature: **tone**:
human-like vs. robotic

93 participants · fantasy-village game

code review

recommends accepting or rejecting a pull request (a Python function)

feature: **length**:
verbose vs. concise

35 participants · software-company game

both games were also played by 50 AI agents each

sanity check: does it measure trust?

1.61× / 1.46× — odds of following advice per SD of accuracy (decision support / code review): more accurate systems are more trusted

no effect — of the round's stakes, so it is not just measuring risk tolerance

r = 0.27 — between behavior and self-report, consistent with prior work

trust bug found: a robotic tone increases trust

1.32×

odds of following the advice with a robotic tone vs. a human-like one, at the same accuracy (p = 0.02)

but people didn't say so

self-reported trust didn't change.
Human-like tone: what people said predicted what they did.
Robotic tone: it didn't.

testing can also show the absence of a bug

length: no trust bug

verbose vs. concise code-review explanations had no effect on behavioral or perceived trust

upper bound of the 95% CI on the odds ratio is 1.6, below the 1.67 minimum effect of interest, so Sam can ship the cheaper model

AI agents: no tone bug

agents' trust tracked accuracy even more strongly than humans' (odds ratio 4.13)

but tone had no effect: the bug affects humans only

trust depends on more than correctness

who is asking

· gender
· for me or for a client ·
· whether the model favors me

how it is shown

· text vs. bars
· a warning label
· explanation design

how it talks

· tone

so: measure behavior, aim for **calibrated** trust,
and test for trust bugs before you ship