Learning with Misspecified Models: the case of overconfidence

Jimena Galindo

September 28, 2023

Overconfidence is Costly

OVERCONFIDENCE: Belief that my type is higher than it truly is ("overestimation" as in Moore and Healy (2008))

Overconfidence is Costly

OVERCONFIDENCE: Belief that my type is higher than it truly is ("overestimation" as in Moore and Healy (2008))

It seems to be persistent in various settings.

- Excess entry of entrepreneurs (Camerer and Lovallo, 1999)
- Suboptimal genetic testing and healthcare (Oster et al. 2013)
- Workers overestimate their productivity (Hoffman and Burks, 2020)

Ultimately it leads to sub-optimal choices

Models of Learning

Some of the features that theory has incorporated to explain overconfidence are:

- Dogmatism
- Paradigm shifts
- Motivated beliefs
- Myopic optimiztion

Four Theories of Misspecified Learning

1. Self-defeating equilibrium (Heidhues et al. (2018)):
    - Bayesian on ω
    - Dogmatic about θ
2. Bayesian Likelihood Ratio test (Schwarstein and Sunderam (2021), Ba, (2022 JMP)) :
    - Bayesian on ω
    - Hypothesis testing on θ
3. Motivated Beliefs or Self-Attribution Bias (Benjamin, 2019):
    - Errors in probabilistic reasoning and judgment biases

An Example

A student has **unknown intrinsic ability** θ^* and chooses a level of effort $e \geq 0$.

Effort and ability are transformed into a noisy output at an exogenous and **unknown rate** ω.

An overconfident student believes he is of type $\hat{\theta} > \theta^*$

And wants to maximize utility

$$y = (\theta^* + e)\omega - \frac{1}{2}e^2 + \varepsilon$$

An Example

A student has **unknown intrinsic ability** θ^* and chooses a level of effort $e \geq 0$.

Effort and ability are transformed into a noisy output at an exogenous and **unknown rate** ω.

An overconfident student believes he is of type $\hat{\theta} > \theta^*$

And wants to maximize utility

$$y = (\theta^* + e)\omega - \frac{1}{2}e^2 + \varepsilon$$

Regardless of their own type, they should choose $e^*(\omega) = \omega$

Learning is Possible

This exercise is repeated for $t = 0, 1, ...$

$$y_t = (\theta^* + e_t)\omega - \frac{1}{2}e_t^2 + \varepsilon_t$$

Note that both parameters are identified in this setting:

- Choosing $\hat{e}$ and $\hat{e} + 1$ over multiple periods allows identification of ω
- Once ω is known, θ can be backed out

How come people don't learn their true type and don't choose the optimal effort?

Mental Models

A prior belief over parameters/states/types and an updating procedure

- Bayesian
- Dogmatic
- Motivated Beliefs/Self-attribution

Research Questions

To what extent do the different theories explain observed behavior?

- Do we observe heterogeneity in the use of mental models?

Is ego-relevance of the type a key feature for the misspecification?

- Can the same theories be used to explain the prevalence of stereotypes?

Road-map

1. Three Theories of Overconfidence
2. Mechanisms and Predictions
3. Unifying Framework
4. Experimental Design
5. Results (coming soon)

The Theories

Settings with two or more unknowns allow for different explanations of the bias:

1. Self-defeating equilibrium (Heidhues et al., 2018):
    - Bayesian on ω
    - Dogmatic about θ

2. Bayesian Likelihood Ratio test (Ba, 2022 JMP):
    - Bayesian on ω
    - Hypothesis testing on θ

3. Self-Serving Attribution Bias with two unknowns (Brunnermeier and Parker, 2005; Coutts et al. 2022wp):
    - Good news are attributed to high θ bad news are attributed to low ω

Unrealistic Expectations and Misguided Learning

(Heidhues, Köszegi, and Strack, 2018)

The Setting

The student's true ability is θ^*, they believe with certainty that it is $\hat{\theta} > \theta^*$.

The rate ω is drawn from density g_0 with $\omega^* = E_{g_0}(\omega)$.

At $t = 0$, the student has the prior g_0.

They correctly choose $e_0 = \omega^*$.

The Setting

The student's true ability is θ^*, they believe with certainty that it is $\hat{\theta} > \theta^*$.

The rate ω is drawn from density g_0 with $\omega^* = E_{g_0}(\omega)$.

At $t = 0$, the student has the prior g_0.

They correctly choose $e_0 = \omega^*$.

Suppose they don't update their beliefs or their choice for a number of periods.

Updating the Beliefs

For their chosen effort ω^*, they observe an average output of

$$y_0 = (\theta^* + \omega^*)\omega^* - \frac{1}{2}(\omega^*)^2$$

But were expecting

$$(\hat{\theta} + \omega^*)\omega^* - \frac{1}{2}(\omega^*)^2 > y_0$$

Updating the Beliefs

For their chosen effort ω^*, they observe an average output of

$$y_0 = (\theta^* + \omega^*)\omega^* - \frac{1}{2}(\omega^*)^2$$

But were expecting

$$(\hat{\theta} + \omega^*)\omega^* - \frac{1}{2}(\omega^*)^2 > y_0$$

So they conclude that ω_1 must be such that:

$$(\hat{\theta} + \omega^*)\omega_1 - \frac{1}{2}(\omega^*)^2 = (\theta^* + \omega^*)\omega^* - \frac{1}{2}(\omega^*)^2$$

Which gives $\omega_1 = \frac{(\theta^*+\omega^*)\omega^*}{(\hat{\theta}+\omega^*)} < \omega^*$

Updating the Beliefs

Updating choices every period (myopically) the belief will drift even further:

A lower choice of e still gives a lower output than expected.

So ω_{t+1} must be lower than they believed in period t.

Prediction: convergence to a self-confirming equilibrium with $\omega_\infty < \omega_1 < \omega^*$.

The result is symmetric for underconfident subjects.

Robust Misspecified Models and Paradigm Shifts
(Ba, 2022 JMP)

The Setting

Same as HKS but with finite Ω and finite A

Now the entrepreneur is willing to switch to an alternative level of ability θ' (assume $\theta' = \theta^*$).

Instead of updating $P[\theta]$ every period, they perform a Bayesian hypothesis test:

Adopt model θ' at time t iff

$$\frac{\ell_t(\theta')}{\ell_t(\hat{\theta})} > \alpha \geq 1$$

Where

$$\ell_t(\theta) := \sum_{\omega} g_0(\omega) \prod_{\tau=0}^{t-1} \pi^{\theta}(y_\tau | a_\tau, \omega)$$

Results

Prediction: Misspecified agents escape the trap as long as their prior is not too "tight" around a self-confirming equilibrium.

Errors in probabilistic reasoning and judgment biases

(Benjamin, 2019)

The Setting

Fixed effort e, $\theta \in \{\theta_H, \theta_L\}$ and $\omega \in \{\omega_H, \omega_L\}$ generate binary signals (**s**/**f**)

After a signal realization m, the agent updates their belief about θ with distortions c_m^θ and c_m^ω, so that:

$$\frac{p_{t+1}[\theta_H]}{p_{t+1}[\theta_L]} = \left(\frac{p[\mathrm{m}|\theta_H]}{p[\mathrm{m}|\theta_L]}\right)^{c_m^\theta} \frac{p_t[\theta_H]}{p_t[\theta_L]}$$

and

$$\frac{p_{t+1}[\omega_H]}{p_{t+1}[\omega_L]} = \left(\frac{p[\mathrm{m}|\omega_H]}{p[\mathrm{m}|\omega_L]}\right)^{c_m^\omega} \frac{p_t[\omega_H]}{p_t[\omega_L]}$$

The agent suffers from self-attribution bias if $c_s^\theta > c_f^\theta$ and $c_s^\omega < c_f^\omega$.

Predictions

Prediction: Even unbiased agents will overweight θ_H after a success and end up being biased.

When $c^\theta = c^\omega = 1$, the updating procedure coincides with the unbiased Bayesian.

The framework does not allow direct comparisons with the other two theories.

A Unifying Framework

Finite type space: $\theta \in \{\theta_H, \theta_M, \theta_L\}$

Finite state space: $\omega \in \{\omega_H, \omega_M, \omega_L\}$ with $p(\omega_k) = 1/3$

Finite action space: $e \in \{e_H, e_M, e_L\}$

Binary signal: Success/Failure with $P\left[Success|e, \omega, \theta\right]$ satisfying the assumptions of HKS

The Data Generating Process

The probability of success is given by:

	ω_H	ω_M	ω_L
e_H	50	20	2
e_M	45	30	7
e_L	40	25	20

θ_L

	ω_H	ω_M	ω_L
e_H	80	50	5
e_M	69	65	30
e_L	65	45	40

θ_M

	ω_H	ω_M	ω_L
e_H	98	65	25
e_M	80	69	35
e_L	75	55	45

θ_H

The Data Generating Process

	ω_H	ω_M	ω_L
e_H	50	20	2
e_M	45	30	7
e_L	40	25	20

θ_L

	ω_H	ω_M	ω_L
e_H	80	50	5
e_M	69	65	30
e_L	65	45	40

θ_M

	ω_H	ω_M	ω_L
e_H	98	65	25
e_M	80	69	35
e_L	75	55	45

θ_H

The Data Generating Process

	ω_H	ω_M	ω_L
e_H	50	20	2
e_M	45	30	7
e_L	40	25	20

θ_L

	ω_H	ω_M	ω_L
e_H	80	50	5
e_M	69	65	30
e_L	65	45	40

θ_M

	ω_H	ω_M	ω_L
e_H	98	65	25
e_M	80	69	35
e_L	75	55	45

θ_H

A Self-Confirming Equilibrium

θ_L

	ω_H	ω_M	ω_L
e_H	50	20	2
e_M	45	30	7
e_L	40	25	20

θ_M

	ω_H	ω_M	ω_L
e_H	80	50	5
e_M	69	65	30
e_L	65	45	40

θ_H

	ω_H	ω_M	ω_L
e_H	98	65	25
e_M	80	69	35
e_L	75	55	45

The Self-Confirming Equilibria

θ_L

	ω_H	ω_M	ω_L
e_H	50	20	2
e_M	45	30	7
e_L	40	25	20

θ_M

	ω_H	ω_M	ω_L
e_H	80	50	5
e_M	69	65	30
e_L	65	45	40

θ_H

	ω_H	ω_M	ω_L
e_H	98	65	25
e_M	80	69	35
e_L	75	55	45

An Example

- True type is θ_M
- True exchange rate is $\omega_M \rightarrow$ The entrepreneur believes it is uniformly distributed

θ_L

	ω_H	ω_M	ω_L
e_H	50	20	2
e_M	45	30	7
e_L	40	25	20

θ_M

	ω_H	ω_M	ω_L
e_H	80	50	5
e_M	69	65	30
e_L	65	45	40

θ_H

	ω_H	ω_M	ω_L
e_H	98	65	25
e_M	80	69	35
e_L	75	55	45

Example: Dogmatic Modeler

- Theory 1: for a student who believes he is θ_H

1. Chooses e_H and is disappointed $\rightarrow$ adjust belief about ω downward
2. Eventually chooses e_M and is disappointed as well $\rightarrow$ adjust belief about ω
3. Eventually chooses e_L and falls into a self-confirming equilibrium

Dogmatic Overconfident

figures2/dogmatic_over_11.png

Figure 1: $\theta^* = \theta_M$, $\hat{\theta} = \theta_H$, $\omega^* = \omega_M$

Example: Likelihood Testing

- Theory 2: for the same initial belief
- Keeping track of the likelihood of each θ

1. Chooses e_H and is disappointed $\rightarrow$ adjust belief about ω downward
2. Eventually chooses e_M and is disappointed as well $\rightarrow$ adjust belief about ω
3. Eventually chooses e_L and falls into a self-confirming equilibrium
4. At some point, the likelihood of θ_M becomes much larger than that of θ_H and the agent updates their belief

Switcher Overconfident

figures2/switcher_over_11.png

Figure 2: $\theta^* = \theta_M$, $\hat{\theta} = \theta_H$, $\omega^* = \omega_M$, $\alpha = 1.1$

Example: Self-Serving Beliefs

- Theory 3: Start with a diffused prior over θ

1. Chooses e that maximizes utility according to priors
2. Success $\rightarrow$ overweight θ_H and underweight ω_H
3. Failure $\rightarrow$ overweight ω_L underweight θ_L
4. Belief on ω deteriorates a lot after failure streaks
5. Belief on θ increases a lot after success streaks

Self-Serving Bias

figures2/self-serving_11.png

The Simulation

figures2/all_11.png

The Experiment:

Part 1: Set Types

- Quiz: Answer as many questions as you can in 2 minutes:
    - Math, Verbal, Pop-Culture, Science, Us Geography, Sports and Video games

- How many questions do you think you answered correctly in each quiz?
    - Bin1, Bin2, Bin3

The Experiment: Ego-relevant condition

Belief updating and effort choice (One topic at a time)

- Choose an effort
- Receive a sample of 10 signal realizations

11 rounds per topic

Eliciting Beliefs?

- $E[\omega]$ is revealed by their choice of effort
- Eliciting beliefs for θ can incentivize learning in a way that is not consistent with the model

Allow them to see the success rate matrix for only one type.

- Track the matrices they choose to see in each round

The Experiment: Stereotype condition

Observe the characteristics of a participant (Gender, US National or not).

- "What score do you think this participant got in the (topic) quiz?"
- Bin1, Bin2, Bin3

Belief updating and effort choice

- Choose an effort
- Receive a signal realization
    - o The DGP is that of the observed participant

11 rounds (per topic/participant)

Screen

figures2/screen1.png

Screen

figures2/screen2.png

Conclusion

What I hope to get from this design:

- A classification of subjects into one of the models based on their behavior
- If subjects are switchers: what is the switching threshold α
- Insight into the role of ego-relevant parameters in belief misspecification

The end

Thank you!