UNIVERSITY OF WARSAW | FACULTY OF ECONOMIC SCIENCES

WORKING PAPERS
No. 14/2026 (508)

KNOW THYSELF: A METHODOLOGICAL MANIFESTO FOR TEACHING MICROECONOMICS THROUGH EPISTEMIC PROVOCATION

TOMASZ KOPCZEWSKI

WARSAW 2026

ISSN 2957-0506

WORKING PAPERS 14/2026 (508)

Know Thyself: A Methodological Manifesto for Teaching Microeconomics Through Epistemic Provocation

Tomasz Kopczewski[1*]

[1] *University of Warsaw, Faculty of Economic Sciences*
[*] *Corresponding author: tkopczewski@wne.uw.edu.pl*

Abstract: This paper documents and formalises the Know Thyself method, a teaching approach developed through more than thirty years of university teaching practice. Its starting diagnosis is that students often learn economic models without experiencing the assumptions that make those models necessary. The method reverses the usual sequence: experience before theory, data as a mirror before abstraction, and questions before answers. Its empirical core is not the experiment narrowly understood, but ad hoc research: classroom experiments, surveys, simulations, valuation tasks, and replication laboratories that make learners' own assumptions visible. Four case studies — expected value, ergodicity, market equilibrium, and the rationality of altruism — illustrate how the method converts declarative knowledge into reflective practice. Artificial intelligence gives the method scale by lowering the cost of surveys, dashboards, simulations, and replication protocols. The paper's practical conclusion is simple: change the order. Ask first. Teach later.

Keywords: economics education, epistemic provocation, experiential learning, replication, expected value, ergodicity, market equilibrium, AI in education, science curiosity

JEL codes: A22, B41, C92, D81, D83

1. The Opening Paradox: Speaking Prose Without Knowing It

> *"Par ma foi! Il y a plus de quarante ans que je dis de la prose sans que j'en susse rien."*
> - Molière, Le Bourgeois Gentilhomme, Act II, Scene 4 (1670)

When Monsieur Jourdain discovers, at the age of forty, that he has been speaking prose all his life without knowing it, his reaction is delight. *Vive la science!* - Long live science! He is grateful, charmed, and immediately eager to put his new knowledge to use in a love letter. The discovery costs him nothing.

The economist who discovers, after thirty years of teaching, that they have been assuming no path dependence all their life without knowing it - that every supply-demand diagram they drew, every expected value problem they assigned, every model of intertemporal choice they transmitted to students contained a hidden claim about the nature of time that they never examined, never named, never offered as a choice - that economist cannot simply shout *Vive la science!* They must ask a harder question: what did I teach, and to whom?

This paper is about that harder question. It is about a method of teaching microeconomics that begins not with theory but with the moment of discovery - the moment a student sees, in their own data, that they do not know what they thought they knew. It is also, unavoidably, about the teachers who design that moment. And it is about what happens when the teacher has never experienced it themselves.

1.1 The Paradox in Four Scenes

Scene one: the expected value experiment. A group of economics students - second-year, introductory microeconomics, all of whom have been taught the expected value formula – is asked to evaluate a series of real-life willingness-to-pay scenarios involving lotteries with varying probabilities of financial or health-related outcomes. The experiment runs on a digital platform. Each decision is timestamped.

The median response time across all probability-weighted scenarios is approximately sixty seconds. The minimum time required to perform even a rudimentary expected value calculation - identify the outcomes, assign probabilities, compute the weighted sum - is two to three minutes. The data are unambiguous: the students did not calculate. They estimated, approximated, or followed some fast heuristic that bypassed the formal tool entirely. When

asked, after the experiment, whether they had used the expected value formula, the majority said yes.

This is not a story about deception. It is not a story about irrationality. It is a story about a tool that was learned without being internalised - a formula transmitted as a symbol to be reproduced on an exam rather than as an instrument to be reached for when a decision is difficult. The gap between the declaration and the clock is not a character flaw. It is the predictable consequence of teaching answers before questions.

Scene two: the risk lecture. An experienced microeconomics professor - publishing record, teaching awards, twenty years in the discipline - faces the same problem every semester when writing exams: how to combine risk and time in a single question without the result collapsing into either a routine calculation or an exercise in advanced microeconomic theory that students dread. The frustration is real. The reflection it triggers is deeper than the exam.

The question arises during a seminar on heterodox economics: what assumption does the standard growth model make about the relationship between cross-sectional averages and individual time paths? The question is not hostile. It is methodological. The answer - that the standard model assumes ergodicity, that it treats a one-shot snapshot of a population as informative about any individual's long-run trajectory, and that this assumption is violated by precisely the path-dependent, compounding dynamics that generate the most consequential economic inequalities - is not obscure. It is available in any careful reading of the literature from Peters (2019) back through Robinson (1962). The professor is not aware of having made this assumption. They have been making it, and teaching it, for twenty years. Their students have been making it without knowing they were. The students' students, in many cases, are making it now.

This is Molière's prose, transmitted across generations. The difference from Jourdain is structural: Jourdain spoke prose in his drawing room, where the consequences were social comedy. The economist assumed no path dependence in their models of wealth distribution, inequality, and policy evaluation, in which the consequences are purely analytical.

*Scene three: the ice cream problem***.** An experienced microeconomics teacher opens the chapter on consumer choice. Again. Every semester, the same diagram: a student - call him Jaś-has 30 zlotys from his mother and must decide between ice cream and chocolate. The preferences are complete and transitive. The budget constraint is binding. The solution is elegant. And every semester, the same discomfort. Not with the mathematics. With the story.

The story says that a human being is someone who allocates resources among goods. Goods are things you are willing to pay for. Introductory microeconomics is, in this sense, a book about choices in a pastry shop. The story does not say that other people exist. It does not say that what you owe to others, or what they owe to you, is part of the choice. It does not say that giving - time, money, attention - might be something a rational agent chooses. The word altruism appears in the standard introductory textbook exactly once across a survey of 20 volumes. The word ethics appears 10 times, always in phrases such as "ethical judgement," never as part of "consumer choice". Ice cream appears 250 times. Homo oeconomicus himself appears once. The teacher knows what this silence produces. Students asked to describe the rational agent at the centre of their models reach for Scrooge. Or more caricatural Scrooge McDuck. Or the calculating sociopath of some recent streaming series. The textbook never said selfishness is rational. It said nothing about selfishness at all. But the student who has spent a semester with ice cream and chocolate has absorbed a story the textbook never explicitly told - that rationality and egoism are the same thing, that to be consistent in one's preferences is to be indifferent to others, that the economics classroom is a place where moral questions do not belong.

The teacher would like to say, "This is wrong." Rationality requires only complete and transitive preferences. Before his transformation, Scrooge was rational in an egoistic way. After his transformation, Scrooge can also be rational in an altruistic way. But in his whole life, he was irrational because he had inconsistent preferences. The content of those preferences - selfish or generous, calculating or caring - is entirely outside the model's scope. The textbook excludes ethics not because ethics is irrational, but because ice cream and chocolate are, after all, so much more *practical*. Closer to reality. Easier to teach. And the exam writes itself.

But saying the rest of it - that altruism can be modelled, that rationality and generosity are not contradictions, that the budget constraint can contain other people - requires tools that do not exist at the introductory level. The curriculum was not designed to hold this argument. There is no chapter for it, no standard exercise, no place in the assessment where it could land. So, the ice cream stays. And the gap between the model and the world it claims to describe stays with it.

Scene four: the teacher who did not believe in the invisible auctioneer. The supply-demand diagram is the most reproduced image in the social sciences. Two lines cross. The crossing point is called equilibrium. Every student who has taken introductory economics has drawn it, been examined on it, and carried it forward as a description of how markets work.

Every semester, the same question arrives - usually from a student who has been paying attention rather than one who has stopped: how do we get there? Not what is the equilibrium. How does the price actually reach it?

The direct answer is uncomfortable. The standard model has no answer. It has a before and an after - a disequilibrium price and an equilibrium price - connected by nothing more than the logic of comparative statics. The ceteris paribus clause freezes time. The model shows two still photographs and asks the student to infer the film. Marshall buried the mechanism of adjustment in the margin. Walras inserted a fictional auctioneer - an omniscient market-maker who calls out prices, collects bids, and announces the equilibrium before any actual trade occurs. No trade happens out of equilibrium. The process is instantaneous. The auctioneer is God.

The teacher knows this is a fiction. There is no auctioneer. In real markets, trades happen continuously, out of equilibrium, at prices that are wrong. The cobweb model - which at least tried to describe the adjustment path - was quietly removed from most curricula sometime in the 1980s. What replaced it was comparative statics: before and after, with the middle omitted. The invisible hand became invisible in a new sense: not present but unexplained, a miracle inserted between two equilibria. The student who learns supply and demand without learning this history does not merely miss some context. They acquire a belief: that markets clear because supply and demand, and that the equilibrium emerges as naturally as water finding its level. They carry this belief into professional life, into policy recommendations, into further teaching. They speak Marshall's prose for forty years without knowing it.

The teacher would like to explain the mechanism. But the mechanism is institutional – it depends on specific trading rules, specific information structures, specific ways in which buyers and sellers find each other and agree on prices. These are not in the standard model. The standard model assumes they are always already satisfied, which is another way of saying it assumes the auctioneer without naming him. And explaining this properly requires more than a lecture. It requires students to be inside a market while it is happening - to feel the pressure of a standing offer, to discover that prices converge not because of mathematics but because of rules.

The teacher has read Vernon Smith's 1962 paper. Has taught it, cited it, assigned it. Has explained, many times, that Smith changed one thing - the trading institution - and produced convergence where Chamberlin had produced chaos. Understands the argument completely.

This semester, for the first time, there will be an experiment. The decision arrives somewhere between the diagram and the next slide, without drama. Not a revelation - more like a concession: the gap between the model and the mechanism has been papered over long enough. The theory says prices should converge. But the theory also said the auctioneer was unnecessary, and then quietly put him back in. The teacher is no longer entirely sure, having taught the model for twenty years without ever testing it, which side of the paper is real. The experiment is scheduled. The students do not yet know what they are about to participate in. Neither, in one important sense, did the teacher believe it would work.

Not in a room of students who had never traded before, who were distracted, who were checking their phones, who had no financial stake in the outcome. The theory said it should work. The intuition - formed by years of teaching the model as an abstract construct, never as a lived event - said it would produce chaos.

The prices converged. Not perfectly, not instantly, but within the range that Smith had documented forty years earlier, in a room not so different. The teacher watched it happen and felt something that, in retrospect, can only be described as the opposite of Jourdain's delight. Jourdain was charmed to discover he had been speaking prose. This was the opposite: the unsettlement of discovering that one had been teaching something one had not believed - and that the thing one had not believed was true.

That unsettlement is the beginning of the Know Thyself method. Not as a biographical curiosity, but as a claim about what kind of knowledge is required to teach economics well. McCloskey (1985) asked whether competitive equilibrium was a unicorn - a creature of extraordinary theoretical beauty and zero confirmed sightings. The question was rhetorical; a provocation directed at the discipline's habit of teaching models as articles of faith rather than empirical claims. But for most economics teachers, the question is not rhetorical at all. It is literal. They have never seen the unicorn. They teach its existence on authority - the authority of textbooks, of their own teachers, of a professional consensus they absorbed without examining.

The experiment changed the relationship to the model. Not the knowledge of it - the result was known before the experiment ran. It changed what could be done with the knowledge: which assumptions could be seen as choices rather than axioms, which questions could be asked that had not been known were askable. This is Paul's (2014) transformative experience in its precise sense: not an experience that adds information to an existing framework, but one that changes the framework itself.

The Know Thyself method is, in part, an attempt to give students a version of that experience - controlled, structured, and pedagogically intentional - before twenty years of teaching have calcified the assumptions they do not know they are making.

1.2 The Structural Problem: Unconscious Transmission

The four scenes above are drawn from a single career. That is, in one sense, their limitation: they are not a sample, not a survey, not a systematic study of what economics teachers experience. They are what one person noticed, over thirty years, in the gap between the models and the rooms.

But the gap is not personal. The textbooks are the same textbooks. The diagrams are the same diagrams. The silence around altruism, around path dependence, around the institutional conditions for equilibrium - these are not quirks of one curriculum or one university. They are structural features of how the discipline transmits itself, reproduced in every introductory course in every country where economics is taught in the standard way.

The frustration in these scenes is probably not unique. The teacher who cannot fit ethics into the consumer choice chapter without losing half the lecture - there are others. The teacher who has explained the invisible auctioneer for twenty years without quite believing in him - there are others. The teacher who discovers, the first time they run a classroom experiment, that the thing they taught as abstract truth is also empirically real - and that the discovery unsettles rather than reassures - there are almost certainly others, and their scenes would be different from these, and would add to the same pattern.

What these four scenes share is not student failure. What they share is a specific mechanism of pedagogical transmission: the model arrives without its story, and the assumption travels without its flag. These four scenes are not the full inventory of what is broken - they are four entry points into a structural problem that has more entry points than any single career can document.

Each of these is a place where the curriculum transmits not just knowledge but the unconscious assumption baked into knowledge - and where the assumption travels without its flag because it was never flagged as an assumption. Because it was only ever treated as a convenience.

Uskali Mäki (1992, 2009) has described this with precision in his account of economic models as isolations: a model is not an approximation of reality but a deliberate excision, setting

aside everything except the mechanism under examination. The ceteris paribus clause is the instrument of this isolation. It does not say that other things are equal in the world. It says: we are setting them aside, in this model, in order to see this one thing clearly. Isolation has costs - every act of isolation excludes something from the analysis. Teaching a model without teaching its isolations is teaching the output while hiding the conditions.

Dani Rodrik (2015) reaches the same diagnosis from a different direction. The core methodological failure of economics education is not that students learn wrong models. It is that they learn to treat one model as the model - to absorb a single analytical frame as a description of reality rather than as one map among many, each appropriate to specific terrain. The skill of a well-trained economist, on Rodrik's account, is the judgment to select the right map. That judgment requires having encountered the model's assumptions as choices: decisions that could have been made differently, that were in fact contested historically, and that produce different results when varied.

Neither Mäki's critique nor Rodrik's is new. The methodological literature has diagnosed the condition repeatedly (Hausman, 1992; Hands, 2001; Backhouse, 2008), and the economics-education literature has made adjacent diagnoses about the profession's training culture (Colander, 2005; Earle, Moran, & Ward-Perkins, 2017). What the literature has not, in the main, produced is a remedy that works inside the standard curriculum - one that does not require a separate optional course in methodology, attended by a self-selecting minority, but that operates within the economics classroom itself, in real time, with ordinary students.

The Know Thyself method is a response to this pattern - not to one teacher's frustration, but to the structural feature that makes the frustration so recognisable. The scenes above are the diagnosis. What follows is the attempt at a remedy.

1.3 Why the Remedy Must Begin with Experience - and Why More Knowledge Is Not Enough

The standard response to the diagnosis above is curricular: add history of economic thought, add methodology, require students to read McCloskey (1985) alongside Varian (2010). This response is not wrong. It is insufficient - and the insufficiency runs deeper than is usually acknowledged.

The information problem. Philip Mirowski (2013) has argued, with characteristic sharpness, that information is not knowledge. The two are not on a continuum where more

information eventually becomes knowledge; they are structurally different things. Information is a signal that can be stored, transmitted, and reproduced without being understood. Knowledge is a capacity to act differently in the world because of what you have encountered. A student who can reproduce the definition of ergodicity on an exam has information. A student who changes how they read a model of wealth distribution because they have understood what ergodicity assumes has knowledge. The standard curriculum produces the first. It has very little mechanism for producing the second.

This distinction matters practically. Adding more content to the economics curriculum - more history, more methodology, more critical readings - increases the information load. It does not, by itself, convert that information into the kind of knowledge that changes how a student uses a model. It may, in fact, increase the problem: a student who has been told more facts about the limitations of economics has more material to deploy in rationalising whatever position they already hold.

The knowledge paradox. This is where the research of Dan Kahan and colleagues delivers its most uncomfortable finding for economics educators. Kahan et al. (2017) set out to test a reassuring hypothesis: that science literacy and cognitive sophistication protect people from politically motivated reasoning - that knowing more makes you more open to evidence that challenges your prior beliefs. The hypothesis is wrong in the direction that matters most for education. Higher science literacy correlates with greater polarisation on contested empirical questions, not less. More analytically sophisticated individuals are better, not worse, at constructing motivated arguments for the position they already hold. Knowledge, without something else, is a tool for defending conclusions rather than for revising them.

The variable that does predict open-minded engagement with disconfirming evidence is not knowledge. It is science curiosity - defined by Kahan et al. (2017) as the disposition to seek out and consume scientific information for the intrinsic pleasure of the encounter, independent of whether that information confirms or challenges existing beliefs. Science curiosity is not the same as science literacy. A highly literate person with low science curiosity uses their knowledge defensively. A person with high science curiosity - even one with lower formal training - engages with anomalies rather than explaining them away.

The implication for economics education is precise and uncomfortable: a curriculum that adds more content without producing science curiosity may be producing more sophisticated defenders of received wisdom, not more reflective practitioners. The student who learns more about the limitations of the expected value model, without ever having experienced the gap

between their declared and actual use of it, has new ammunition for a sophisticated critique - and no changed relationship to the tool itself.

The curiosity mechanism. How is science curiosity produced? Kahan et al. do not answer this question directly - their paper identifies the variable and its effects, not its origins. But the cognitive science of curiosity offers a partial answer. Golman and Loewenstein (2018) characterise curiosity as a response to a specific kind of information gap: the awareness that there is something one does not know, combined with the belief that finding out is possible and that the answer will matter. Curiosity is not a general disposition toward learning. It is triggered by a specific experienced incompleteness - the felt presence of a question without an answer.

This is the mechanism the Know Thyself method is designed to activate. The ad hoc research does not teach students about the expected value formula. It creates the conditions under which students discover, in their own timestamped data, that they did not use the formula when they should have. The gap between their declaration and the clock is not communicated to them as a fact. It is experienced as a question that belongs to them personally: why did I not calculate? what did I actually do? and what should I have done? That question - arising from their own data, about their own behaviour - is the information gap that Golman and Loewenstein describe. It is also, if the method works, the beginning of science curiosity directed at the tool the experiment was designed to introduce.

The sequence matters. Robin Hertwig and Ido Erev (2004) document this with precision in their account of the *description-experience gap*: people who are told the probability of an outcome respond to it differently from people who have experienced outcomes drawn from that distribution. Described risk and experienced risk activate different cognitive mechanisms, produce different judgements, and lead to different decisions - even when the objective probabilities are identical. Description tells you a fact about the world. Experience changes your relationship to it.

The implication is direct. Telling a student that ergodicity is an assumption is a description. Showing a student, in their own data, that their expected wealth after a sequence of multiplicative shocks diverges from the arithmetic average they computed - that is, experience. The first produces a fact to be reproduced on an exam, and possibly deployed in a sophisticated critique. The second produces a question the student cannot un-ask, and a curiosity that the critique alone cannot generate.

Laurie Ann Paul (2014) has theorised the strongest version of this distinction under the name of *transformative experience*: an experience that changes not just what you know but who you are as a knower - your preferences, your categories, your sense of what questions are worth asking. Paul's account is primarily philosophical, but its pedagogical implication is practical: if the goal of economics education is to produce economists who can use tools competently - and, more importantly, who know when they are failing to use them, and who remain curious about the gap - then the educational sequence must include experiences that transform the student's relationship to the tools, not merely their declarative knowledge of them.

This is the argument that motivates the Know Thyself method. Not that experience is pedagogically pleasant, or that active learning improves retention. The argument is epistemological and, in light of Kahan et al., political in a broader sense: a curriculum that produces science curiosity - not just science literacy - is a curriculum that produces citizens less susceptible to the motivated reasoning that polarises democratic societies. The economics classroom is not insulated from this problem. It may, if it continues to transmit tools without their stories and models without their assumptions, be actively contributing to it.

1.4 The Jourdain Problem, Revisited

Monsieur Jourdain's discovery was delightful because the stakes were low. He had been speaking prose in his drawing room. Nobody was harmed by his ignorance. When he learned the word, he was charmed. The economics teacher who discovers they have been assuming ergodicity faces a harder reckoning. The stakes are not low. They have been producing graduates who model wealth distribution, advise pension funds, design social insurance systems, and teach the next generation of students - all on the basis of an assumption they did not know they were making. The discovery is not an occasion for delight. It is an occasion for a question that most curricula are not designed to address: what did I actually teach, and what should I do differently now?

The Know Thyself method does not resolve this reckoning for the teacher. What it does is create the conditions under which the student - and, if the teacher uses it properly, the teacher - encounters their own assumption before it has been transmitted for forty years. The ad hoc research situation comes before the theory. The student sees themselves in the data before they are told what the data mean. The question arrives before the answer.

Jourdain, after his discovery, immediately wanted to write the love letter correctly. The desire to act on the new knowledge was instant. The Know Thyself method bets on

the same mechanism: that a student who has experienced the failure of a tool they thought they had will reach for the correct tool with a motivation that no amount of lecturing can produce.

The method does not begin with what students do not know. It begins with what they discover, in their own data, that they were doing without knowing it.

2. The Hidden Ontology of Economic Models: Every Model Is a Story About What Kind of Creature You Are

"The king is naked."

- Hans Christian Andersen, The Emperor's New Clothes (1837)

The child in Andersen's story does not possess superior knowledge. They possess inferior socialisation. They have not yet learned that the correct response to an emperor parading in invisible clothes is to describe the clothes. They say what they see. The courtiers - educated, experienced, professionally competent - say what they are supposed to say.

Economics education, at its worst, is a training in courtier behaviour. Students learn to describe the clothes. They learn the vocabulary of rational agents, optimising behaviour, equilibrium outcomes, and market efficiency - and they deploy it fluently, professionally, without examining whether the clothes are there. The Know Thyself method is, in part, an attempt to recover the child's directness: to ask, before teaching the model, what the model is actually claiming about the humans it describes. The answer, in almost every case, is: more than the model admits.

2.1 The Ontological Layer That Textbooks Do Not Teach

Every economic model contains an unstated story about what kind of creature a human being is. This story is not presented as a story. It is presented as an assumption - a technical simplification, a modelling convenience, a starting point that everyone agrees to for purposes of tractability, with Friedman (1953) as the canonical methodological defence of this move. What it actually is, in Mäki's (1992) precise sense, is an isolation: a deliberate exclusion of everything about human nature that the model cannot handle, in order to make the mechanism the model can handle analytically visible.

The isolation is not a flaw. It is the method. The flaw is in the transmission: the isolation travels from paper to textbook to lecture to student without its flag. The student does not learn that the model has set aside a specific feature of human behaviour. They learn the model as if

the feature had never existed. The assumption becomes invisible not because it was hidden but because no one mentioned it was there.

Gibbard and Varian (1978) were precise about what this means: economic models are caricatures - deliberate distortions that exaggerate certain features to make them analytically tractable. A caricature is not a lie. It is a selective truth. The problem arises when the caricature is mistaken for a portrait - when the selective truth is transmitted as the whole truth, and students spend careers applying a caricature to a world that contains all the features the caricature omitted.

Robert Shiller (2017) has argued that economics abandoned narrative at precisely the moment it became most rigorous - that the mathematisation of the discipline produced tools of great precision and stories of great poverty, and that the poverty of the stories has costs that the precision of the tools cannot compensate. Blaug (2003) described the same historical turn as the formalist revolution of the 1950s. The supply-demand diagram is an instrument of extraordinary analytical power. It is also a story about frictionless exchange, stable preferences, and institutional neutrality that most of its users have never examined and most of its teachers have never named; even the stability of preferences is not an innocuous assumption (Ariely, Loewenstein, & Prelec, 2003).

We propose to name it. Three examples - each from a different part of the standard curriculum, each carrying a hidden ontological claim that the curriculum transmits without acknowledgement.

2.2 Pascal's Wager and the Permission to Calculate

The expected value formula is taught in introductory statistics and economics courses as a computational tool: multiply each outcome by its probability, sum the products, obtain the expected value. The formula is correct. What the formula does not contain - what no textbook mentions in the chapter where the formula appears - is the two-century philosophical struggle required before anyone was permitted to write it down.

Ian Hacking (1975) reconstructs this history with a precision that should be embarrassing to every economist who has taught the formula without it. Before the mid-seventeenth century, the idea of calculating the expected outcome of a risky event was not merely unusual. It was, in the dominant ontological framework of European civilisation, *illegitimate*. Outcomes were determined by fate, providence, or divine will. To calculate the probability of an outcome was

to presume to know what God had not revealed. The merchants and insurers of the thirteenth and fourteenth centuries who developed the practical arithmetic of risk did so in a legal and moral grey zone, using instruments – commenda contracts, marine insurance policies – that the Church regarded with suspicion precisely because they involved quantifying what was supposed to be unquantifiable.

What changed was not the mathematics. The mathematics of probability was available, in various forms, before Pascal. What changed was the ontological permission. Blaise Pascal's wager (~1660) - the argument that one should calculate the expected utility of believing in God, weigh it against the expected utility of disbelief, and choose accordingly - was scandalous not because it was mathematically sophisticated but because it applied the logic of commercial calculation to the question of salvation. If you could calculate whether to believe in God, you could calculate anything. The sacred and the profane had been separated for a millennium. Pascal's wager dissolved the boundary.

This is the revolution that made the expected value formula possible as a habit of mind, not just as a mathematical operation. Before Pascal, the formula existed as arithmetic. After Pascal, it existed as a permission - a socially and ontologically sanctioned way of approaching decisions under uncertainty. A student who learns the formula without this history has learned an answer without a question. They have been handed a tool whose handle carries the inscription: this is how rational people think about risk - without being told that the definition of rational, in this context, was the outcome of a specific theological dispute in seventeenth-century France.

The pedagogical consequence is visible in the data. Students who have been taught the formula, and who can reproduce it correctly on an exam, do not reach for it when facing a real decision under uncertainty. The median response time in our experiments - sixty seconds for decisions requiring three minutes of calculation - documents the gap between declarative and procedural knowledge with a precision that no survey instrument could replicate. The formula was learned. The habit of mind was not transmitted, because the habit of mind requires understanding why the formula is the appropriate response to uncertainty - and that understanding requires the history that the curriculum omitted.

The hidden ontology of expected value: you are a creature for whom it is legitimate, rational, and indeed obligatory to calculate the probability of outcomes before you act. This is not obvious. It took two centuries of theological, philosophical, and commercial struggle to establish it.

2.3 Robinson's Arrow and the Frozen Clock

Joan Robinson (1962) identified what she called the most dangerous assumption in economics with a directness that the profession has spent sixty years not quite absorbing: the assumption that time has an arrow, but models do not.

In everyday life, events occur sequentially. Yesterday's loss cannot be recovered by today's gain - not because the mathematics forbids it, but because you are not the same person, in the same position, with the same options, that you were yesterday. Path matters. History accumulates. The sequence of outcomes, not just their average, determines where you end up. This is what physicists mean by ergodicity: a system is ergodic if time averages and ensemble averages coincide - if the average outcome over many periods for one agent equals the average outcome at one moment across many agents. Many physical systems have this property. Human economic lives, almost universally, do not.

The standard microeconomic model assumes ergodicity without naming it. The expected utility framework treats a sequence of risky choices as equivalent to a single choice evaluated at the expected value of the sequence. The representative agent in a growth model accumulates wealth along a path that is, in effect, the average of all possible paths - not the specific path that history, chance, and compounding have produced for any actual individual. The model freezes time in the precise sense that Robinson identified: it removes the arrow, the accumulation, the irreversibility that make economic life what it is.

Ole Peters (2019) has formalised Robinson's intuition into a precise mathematical claim: that the standard approach in economics uses ensemble averages where time averages are the appropriate measure, and that this substitution produces systematically wrong predictions about individual behaviour under risk - predictions that have been misinterpreted as evidence of irrationality when they are in fact evidence of rationality in a non-ergodic world.

The hidden ontology of the standard model of choice: you are a creature who can be represented as the average of all possible versions of yourself at a single moment in time. You are not a being who lives through time, accumulates history, and faces decisions whose consequences compound. You are a statistical ensemble.

2.4 Marshall's Buried Conditions

Alfred Marshall (1890) gave economics its most reproduced image: two curves intersecting at a point called equilibrium. The image is so familiar, so deeply embedded in the professional training of economists, that it has acquired the epistemological status of a natural law. Students encounter it within weeks of beginning their training and carry it, largely unexamined, for the rest of their careers.

What Marshall's image does not show - what the standard presentation has systematically omitted since the diagram entered the introductory curriculum - is the institutional machinery required to produce the equilibrium it depicts. The Marshallian cross assumes that a specific mechanism exists through which dispersed private information - what each buyer is willing to pay, what each seller is willing to accept - is aggregated into a public price. It buries this assumption in the *ceteris paribus* clause and proceeds.

McCloskey (1985) asked whether competitive equilibrium was a unicorn. The question pointed to a genuine epistemological gap: between a model with impeccable internal validity - given its assumptions, the equilibrium follows with mathematical necessity - and the external validity question the model cannot answer from within itself: has anyone ever actually observed competitive equilibrium emerge in a real market?

Chamberlin (1948) looked. He put students in a room with private valuations and set them loose to trade. The equilibrium did not appear. He concluded that equilibrium theory was wrong. He had not noticed that he was testing the institution as much as the theory - that bilateral search, which was his trading mechanism, does not provide the informational conditions the Marshallian model requires.

Vernon Smith (1962) changed one thing: the institution. He replaced bilateral search with a continuous double auction - a centralised order book in which all bids and asks are publicly visible and trades occur when a bid meets a standing ask. The equilibrium appeared, reliably, within a few trading periods, with ordinary students who had no financial expertise and no knowledge of the theory. Smith had not found a population of more rational agents. He had found an institution that did the work that Marshall's ceteris paribus had silently assumed.

The hidden ontology of the Marshallian cross: you are a creature who trades in markets where the institutional conditions for equilibrium are always already satisfied. The model does not examine whether you are. It assumes you are, and proceeds.

2.5 The Selfish Axiom: Atomism, the Banishment of Kant, and the Slow Return of Moral Sentiments

The three hidden ontologies described in the preceding sections each conceal a specific assumption about the human being inside the model. The fourth is different in kind: it is not a technical assumption buried in a mathematical structure but a foundational claim about human nature that entered the model as a technical convenience and was never asked to leave. The problem is not that homo oeconomicus is sometimes empirically wrong. The problem is that the entire architecture of the introductory curriculum rests on an atomistic ontology - and that this ontology shapes everything that follows from it, including what kinds of ethical questions can be asked and which cannot.

Move one: the behavioural appendix and its schizophrenia. Something changed in introductory economics textbooks sometime in the 1990s. A new chapter appeared behavioral economics - usually near the end, often optional - describing systematic deviations from rational choice: loss aversion, the endowment effect, hyperbolic discounting, the framing effect (Kahneman, Knetsch, & Thaler, 1990). It cited Kahneman and Tversky. It acknowledged that real human beings deviate systematically from the standard model's predictions.

The addition was well-intentioned. It was also, in a precise sense, schizophrenic - a term used here not loosely but in its clinical sense of a split between two incompatible frames that coexist without resolution (Camerer, Loewenstein, & Rabin, 2004; Thaler & Sunstein, 2008). For twenty chapters the textbook had built homo oeconomicus as the foundation of all analysis. Then, in the final chapter, it announced that people are not like this. The diagnosis of irrationality depends entirely on the model it critiques - there is no bias without a benchmark, no deviation without a standard. The student absorbs a curious conclusion: the model is wrong about what people do, but right about what they should do. Homo oeconomicus survives as a normative ideal even after his descriptive credibility has been withdrawn. Behavioural economics, framed as a corrective appendix, preserves the architecture it appears to challenge.

Move two: the deeper problem is atomism. The schizophrenia points to something the behavioural revolution did not address - and by its framing made harder to see. Both the standard model and its behavioural supplement share a more fundamental assumption: the individual agent, reasoning alone. The standard agent has perfect cognitive capacities and applies them in isolation. The behavioural agent has imperfect cognitive capacities and applies them in isolation. In neither case does reasoning have a social dimension. In neither

case do other people enter as partners in the cognitive process rather than as objects in the environment. The correction is applied to the atom. The atom itself is never questioned.

This is the atomistic ontology. It is prior to the question of rationality or irrationality. It is the assumption that human cognition, decision-making, and preference formation are fundamentally individual processes - that what a person reasons, chooses, and values is theirs alone, shaped by their own cognitive capacities and their own history, independent of the ongoing cognitive and social interaction with others that constitutes most of actual human thought.

Move three: three counter-traditions that the textbook ignores. Vernon Smith (2003, 2008) names the problem directly. Both standard and behavioural economics study what he calls constructivist rationality: the individual mind applying its own capacities to problems it faces alone. What neither addresses is ecological rationality: the intelligence that emerges from the interaction of agents with each other and with the institutions, norms, and practices that structure their environment. The rationality is in the system, not the atom. Smith's own experimental work demonstrated this: markets converge to efficient outcomes through mechanisms that no individual fully understands and that no individual could replicate alone.

Mercier and Sperber (2017) extend the argument into cognitive science. Reasoning, they demonstrate, did not evolve primarily as a tool for individual problem-solving. It evolved as a tool for social persuasion and collective justification - for arguing with others, evaluating others' arguments, reaching conclusions that no individual could reach alone. The isolated reasoner of both standard and behavioural economics is not merely an idealisation. It is a misidentification of what reasoning is for.

Gigerenzer (2008) and Gigerenzer, Hertwig, and Pachur (2011) reach a parallel conclusion from the study of heuristics. The fast-and-frugal shortcuts that behavioural economics identifies as biases are, in many environments, ecologically rational - adapted to the structure of the world in which human cognition actually operates. The benchmark of expected utility maximisation against which they are judged as failures is itself an ecological misfit: a standard appropriate to a world of known probabilities and stable preferences that does not describe the environment in which human beings actually make decisions.

All three traditions converge on the same picture: the human agent as fundamentally embedded - in social relationships, in institutional structures, in a collective knowledge that no individual possesses alone. This is not the atom of the standard model. It is not the biased atom

of behavioural economics. It is a different kind of creature entirely, and the introductory textbook has no chapter for it.

Move four: the technical reason, and its ethical consequence. The technical reason for the atom is Plott (1973): allow preferences to depend on the preferences or consumption of others, and the standard conditions for equilibrium existence and stability collapse. Social learning - the process by which agents change their beliefs and preferences through interaction with others - violates both the independence condition and the path independence requirement simultaneously. It cannot be contained within the standard framework without restructuring it from the ground up. The atom is not a moral choice. It is a condition of tractability. It was never labelled as such.

The ethical consequence follows directly. A framework built around the independent, ahistorical atom can accommodate only one kind of ethical reasoning: the kind that aggregates individual utilities and maximises their sum. Utilitarianism fits because it preserves the optimisation structure. The ethical question becomes a maximisation problem with a social objective function. The tools remain the same.

Kant cannot be accommodated. Not because his framework is less rigorous, but because it presupposes precisely what the model excludes: an agent who is constitutively related to others, whose obligations are binding prior to and independent of consequences. Act only according to that maxim by which you can at the same time will that it should become a universal law is not a calculation (Kant, 1785/1997). It makes sense only for an agent embedded in a shared moral world - not floating free as an atom who happens to interact with other atoms in markets. Treat humanity always as an end, never merely as a means is a constraint on optimisation, not a form of it. The model was not constructed in a way that makes Kant's answers wrong. It was constructed in a way that makes his questions unaskable.

Move five: what the textbook does not say, and what the student therefore cannot ask. Adam Smith's Theory of Moral Sentiments (1759) was the companion volume to The Wealth of Nations - the account of the social and moral sentiments that make commercial society possible and constrain its pathologies. For most of the twentieth century the two books were treated as contradictory: the moral sentiments were set aside as prescientific, and the invisible hand was stripped of the sympathetic framework in which Smith had embedded it. The return of moral sentiments to economics - through experimental work on fairness, through the literature on social norms, through growing interest in non-consequentialist ethics - is real

but partial. It occurs at the margins, in specialised journals, in optional seminar courses. It does not appear in the introductory textbook.

The student who completes introductory economics has been given a powerful analytical toolkit and a picture of human beings as isolated optimisers whose moral universe contains only the question of how much. They have not been given the tools to ask whether the isolation is real, whether the preferences are truly independent, whether the obligations that bind them to others are of a kind that no utility function can represent, or whether the path that led here matters.

The curriculum did not intend this. It arrived here through a series of technical conveniences - Plott's independence condition, the ergodicity assumption, the exclusion of path dependence - each of which was adopted for reasons of tractability and none of which was presented as a choice about what kind of creature a human being is. But technical conveniences, accumulated without acknowledgement, become ontologies. And ontologies, transmitted without examination, become the unquestioned background of everything that follows.

The hidden ontology: you are an atom. Your reasoning is your own. Your obligations reduce to calculations. The community of others in which you live and think and owe things to people - that is not in the model.

2.6 The Ontological Turn as Pedagogical Method

These three examples are not selected to attack the models. The expected value formula is an indispensable tool. The ergodic framework has genuine domains of valid application. The Marshallian cross is, as McCloskey would say, a magnificent instrument - provided its conditions are met. The critique is not of the models. It is of the transmission: the habit of teaching the output while hiding the conditions, of transmitting the formula while omitting the permission, of drawing the diagram while burying the institution.

The Know Thyself method proposes a specific corrective. Before teaching the model, ask the ontological question: what kind of creature does this model assume the student to be? Not as a rhetorical gesture, but as a genuine inquiry that precedes the technical introduction. The student who has been asked this question - and who has, through the ad hoc experiment, already experienced one answer to it that surprised them - arrives at the model's formal apparatus with a different orientation than the student who has been handed the formula on slide three.

They arrive as a reader rather than a recipient. They arrive with a question the model may answer, rather than with an answer the model has pre-installed. They arrive, in Kahan's terms, curious - not merely informed.

This is not a claim that every economics course should begin with the history of probability theory or a seminar on Mäki's philosophy of science. The ontological turn does not require an additional course. It requires a different first question - asked in the first session, before the first formula, before the first diagram. The question is simple, and it is ancient: *Know thyself.*

Not as an injunction to introspection. As a methodological demand: before you learn what this model predicts about human behaviour, find out - in your own data, from your own choices - what kind of human being you actually are. Then see whether the model's story fits.

The model is not wrong until you have checked it against yourself. And you cannot check it against yourself until you have seen yourself in the data.

3. The Know Thyself Method: Mechanism and Five Pillars

> *"An unexamined life is not worth living."*
> - Socrates, in Plato, Apology (399 BCE)

Socrates did not lecture. He asked questions until his interlocutor discovered what they did not know. The discovery was never pleasant - it was, by design, unsettling. And it was never about abstract knowledge. It was always about the specific person in front of him: what do you believe, what do you do, and do those two things cohere? The method was personal before it was philosophical.

The Know Thyself method inherits this structure. It is not named after Socrates as a rhetorical gesture. It is named after him because the pedagogical sequence is genuinely Socratic: the ad hoc research situation comes before the theory, the personal data comes before the general principle, and the question the student cannot un-ask arrives before the answer they were going to be given anyway. The difference is scale - Socrates could manage one conversation at a time - and instrumentation. The platform, the nickname, the auto-generated report, the flexdashboard: these are the tools that make the Socratic provocation reproducible at the level of a course, a semester, a discipline.

But the method is not a technique. Techniques can be adopted without understanding their rationale. The Know Thyself method requires the teacher to have understood, at the level of their own experience, what it means to see themselves in data - which is why Scene Four in the previous section was not autobiography but argument. A teacher who has never experienced the unsettlement of finding their own assumption in their own results cannot design the conditions under which a student will experience it. The method begins with the teacher before it begins with the student.

3.1 The Inversion That Defines the Method

Standard economics pedagogy follows a sequence that has become so normalised it is no longer experienced as a choice:

Theory → Example → Exercise

The theory is presented first, as a given. The example illustrates the theory. The exercise tests whether the student can apply the theory to a new case. The student's own experience, intuitions, and behaviour are either irrelevant to this sequence or appear only as illustrations of the theory's predictions, which the student is then expected to confirm. The Know Thyself method inverts this sequence:

Experience → Mirror in Data → History → Theory → Ethics

The student acts before they know what they are supposed to do. They see themselves in the data before they know what the data are supposed to show. They encounter the historical and philosophical struggle that produced the theoretical tool before they are asked to use it. And the ethical question - what should I have done, and what does this imply about how I want to act? - arrives at the end as a genuine question, not as a pedagogical formality.

This inversion is not cosmetic. It changes the epistemological status of every element in the sequence. Theory, in the standard sequence, is an authority to be absorbed. In the Know Thyself sequence, it is a proposed answer to a question the student has already asked - a question that arose from their own data, and that they are therefore motivated to resolve. The difference in motivation is not a matter of engagement or enthusiasm, though those may follow. It is a matter of what kind of knowledge is being produced: declarative knowledge that can be reproduced on an exam, or procedural knowledge that changes how the student acts the next time the situation arises.

Hertwig and Erev's (2004) description-experience gap explains why the inversion works at the cognitive level. But the deeper rationale is epistemological: the student who has experienced the question before receiving the answer has a different relationship to the answer than the student who received the answer first. The first student *owns* the question. The second student has been given an answer to a question they never asked.

3.2 Five Pillars

The method is built on five structural elements. Each is necessary. None is sufficient alone. Together they constitute a system in which each element reinforces the others.

Pillar One: The Hidden Ontology as Entry Point. Before the model is introduced, the teacher asks the ontological question: *what kind of creature does this model assume the student to be?* This is not a warm-up exercise. It is the first act of the method - the moment at which the model's hidden assumptions are named before the model is transmitted. As established in Section 2, every economic model contains an unstated story about human nature that travels invisibly from teacher to student because no one flags it as a choice. The first pillar flags it.

In practice, this means beginning each topic not with the formula or the diagram but with a question about the person. Before expected value: *when you face a decision with uncertain outcomes, what do you actually do?* Before the supply-demand model: *have you ever been in a market where prices emerged from the interaction of buyers and sellers - what happened?* Before ergodicity: *does the average outcome across a population at one moment tell you anything reliable about what will happen to you personally over time?* These questions are not rhetorical. They are genuine empirical inquiries that will be answered - partially, provisionally - by the experiment that follows. *Key sentence: ask what the model assumes about humans before you teach the model.*

Pillar Two: The Reversed Order of Knowledge. The description-experience gap (Hertwig & Erev, 2004) establishes that described and experienced risk activate different cognitive mechanisms. The Know Thyself method operationalises this finding into a pedagogical principle: the student must experience the phenomenon before the theory that explains it is introduced.

This is the most consequential departure from standard pedagogy, and the one most likely to generate resistance from teachers who are accustomed to presenting theory first.

The resistance is understandable: it feels disorderly to put students in an experimental situation before they know what they are supposed to observe. It also feels inefficient: why not simply tell them the result and save the time?

The answer is in the data. A student who has been told that people systematically underweight low-probability events has a piece of information. A student who has just seen, in their own timestamped responses, that their median decision time was sixty seconds when the minimum calculation time was three minutes has an experience - and a question. The information can be forgotten. The question cannot be un-asked.

The sequence is therefore fixed by the method's logic, not by pedagogical preference: experience precedes mirror, mirror precedes history, history precedes theory. Kolb's (1984) experiential learning cycle recognises the same structure - concrete experience, reflective observation, abstract conceptualisation, active experimentation - but the Know Thyself method adds a specific element that Kolb does not foreground: the *mirror*. The student does not merely have an experience and reflect on it. They see themselves in data, specifically, numerically, with a nickname that allows individual identification within aggregate results. The reflection is anchored in personal evidence, not general impression. *Key sentence: experience first, mirror in data second, theory third - in that order, always.*

Pillar Three: Ad Hoc Research as Epistemic Provocation. The ad hoc research situation at the centre of each Know Thyself session is not necessarily a scientific experiment. This distinction must be stated explicitly - in every paper that describes the method, and to students in the session itself. It may take the form of a classroom experiment, a survey, a valuation task, a simulation, a diagnostic exercise, or a replication laboratory. In each case, its function is the same: it creates the conditions under which the participant's own behaviour, judgement, or reconstruction becomes evidence about the question the session will address.

The difference from a behavioural experiment matters methodologically and pedagogically. A behavioural experiment is designed to test a hypothesis about a population, under controlled conditions, with sufficient statistical power to support generalisation. An epistemic provocation is different. It is designed to produce an experience, response, trace, or reconstruction that opens a specific question for the participant. Its function is not to generate publishable data. Its function is to close the description-experience gap.

In practice, these ad hoc research situations may be run through LabSee, survey platforms, classroom tasks, simulations, AI-assisted dashboards, or replication laboratories. Some

generate behavioural data; others generate valuations, classifications, written answers, simulated trajectories, or reconstructed models. What they share is the same sequence: first the participant acts, answers, chooses, simulates, or reconstructs; then the report becomes a mirror; only afterwards does theory enter.

Three design principles govern every ad hoc research situation:

> First: the participant should not know in advance what theoretical point the task is designed to reveal. Prior knowledge of the theoretical prediction converts the experience into a test of the theory rather than a genuine encounter with the phenomenon.
>
> Second: the participant should feel that they are making a meaningful judgement, decision, reconstruction, or interpretation - not merely performing for the teacher. This requires ecological validity: the scenarios must be recognisable, the stakes or questions must feel consequential, and the task must not signal its own artificiality.
>
> Third: the mirror comes after the action. The report, dashboard, or debriefing is generated only after the participant can no longer change the original response. The participant acts in ignorance of the aggregate. They see the aggregate only when their own trace has already been produced.
>
> This sequencing is not incidental. It is the mechanism. *Key sentence: not a scientific study - an epistemic provocation that produces the question the theory will answer.*

Know Thyself is experimental in spirit, but it is not limited to experiments in form. Its basic unit is ad hoc epistemic research: a designed situation in which participants generate evidence about themselves, a model, or an article before theory tells them what that evidence means.

Pillar Four: The Nickname as Sociotechnique

The nickname is not a privacy feature. It is the mechanism of the method. When the report is generated, each student can locate their own result within the group distribution by finding their self-chosen nickname. They can see, in a single visualisation, where they stood relative to the class - whether they were in the majority or the minority, whether their response time was typical or outlying, whether their choices were consistent or scattered. They can see this without anyone else knowing which result is theirs.

This combination - personal visibility with social anonymity - produces a specific psychological condition that is difficult to replicate by other means. The student sees themselves in the data. They cannot deny the evidence, because it is their own, identified by their own name. They are not being shown a statistic about other people. They are being shown a fact about themselves.

At the same time, the anonymity removes the social cost of the discovery. A student who discovers, in a public setting, that they failed to use a tool they claimed to know would experience shame, defensiveness, or strategic silence. The nickname protects against this. The discovery is personal and private, even in a room of a hundred people. The student can be unsettled without being exposed.

This is what the method means by *sociotechnique*: a technical design that produces a specific social and psychological condition - in this case, the condition under which self-knowledge is possible without social threat. The nickname creates the space in which the student can, without embarrassment, ask: was my behaviour typical? was it rational? was it what I would have chosen if I had thought about it differently? These are Socratic questions. The nickname is the technical condition that makes them askable. *Key sentence: the nickname allows the student to find themselves in the data while remaining anonymous to everyone else - this is not a technical feature, it is the method's mechanism.*

Pillar Five: Bidirectional Knowledge Flow. The standard pedagogical relationship is unidirectional: knowledge flows from teacher to student. The student's role is to receive, process, and reproduce. Their individual characteristics - their specific errors, their heterogeneous responses, their particular confusions - are noise to be managed, not signal to be studied.

The Know Thyself method reverses this. The data generated by students in the ad hoc research situation become material for the class itself. The heterogeneity of the group - the fact that some students calculated and some did not, that some were above the median response time and some below, that some were risk-seeking and some risk-averse - is not a pedagogical problem to be averaged away. It is simultaneously the subject of the session and the evidence that the session uses.

This bidirectionality has two consequences. The first is motivational: students who know that their responses will be analysed collectively, and that the analysis will be the content of the class, are participants in a genuine inquiry rather than an audience for a predetermined

lesson. Their data matter. The second consequence is epistemological, and it connects the method to the broader question of what economics education is for. The heterogeneity that the experiment reveals is itself an economic phenomenon of the first order. Why do some people use expected value and others do not? Why do some converge to the equilibrium price and others do not? A curriculum that treats student heterogeneity as noise has decided, in advance, that the theory explains everyone equally well - and that deviations are errors to be corrected rather than evidence to be explained. The Know Thyself method does not make this decision in advance. *Key sentence: the group's heterogeneity is simultaneously the pedagogical problem and the object of study - this is rare, and it is important.*

3.3 The Session Sequence

Each Know Thyself session follows a five-step sequence that implements the five pillars in order. The steps are not interchangeable. The logic of the method depends on their sequence.

Step 0 - The hidden story (Pillar One). The teacher names the model's hidden ontological claim as a question. What does this model assume about you? The question is left open.

Step 1 - Ad hoc research (Pillar Three). The task runs: students act, answer, choose, simulate, classify, or reconstruct without knowing exactly what theoretical point is being tested. Responses or traces are recorded. No results are shown during the task.

Step 2 - Mirror in data (Pillars Two and Four). The report is generated. Students locate themselves using their nickname. The discussion begins with the data: what do you see? does it surprise you? what does it say about what you actually did?

Step 3 - History and theory (Pillar Two, continued). The theoretical framework is introduced as an answer to a question the student has already asked. Pascal's ontological permission, Robinson's frozen clock, Smith's institutional breakthrough: the history explains why the formula or the model exists, and why the student's own behaviour either confirms or departs from its predictions.

Step 4 - Ethical question (Pillar Five). The session closes not with a summary of results but with a question about implications. Having seen what you actually do, and having understood what the model says you should do, what kind of agent do you want to be? This step is optional in shorter sessions but is the natural endpoint of the method's logic.

3.4 What Know Thyself Is Not

The method borrows elements from several established pedagogical traditions. The borrowings are deliberate and acknowledged. What the method does with those elements is not what those traditions do with them, and the differences matter.

~~Not Kolb's experiential learning cycle.~~ Kolb (1984) identifies the correct sequence but does not specify what the experience must contain. The Know Thyself method requires a specific kind of experience: one in which the student's own behaviour becomes evidence about the theoretical question. The experience is not incidental to the theory; it is designed to produce the question the theory will answer.

~~Not flipped classroom.~~ The flipped classroom moves content delivery outside the room but keeps the epistemological order intact: theory before exercise. The Know Thyself method inverts the epistemological order. A flipped classroom that assigns a reading on expected value before the experiment is not a Know Thyself session.

~~Not narrative economics.~~ Shiller (2017) argues that narratives drive economic phenomena. The Know Thyself method uses narrative in a specific direction: to locate the ontological commitments embedded in economic models - to explain why the formula became thinkable, not to explain why markets move.

~~Not generic constructivism.~~ Constructivist pedagogy holds that students learn by building on prior knowledge. The Know Thyself method adds a specific claim: the prior knowledge most relevant to economics education is not the student's existing knowledge of economics, but their knowledge of themselves - their own decision patterns, their own response times, their own departures from the models they are about to learn.

What the method uniquely combines is: the personal data mirror (nickname + report), the ontological question as entry point, the historical grounding of every theoretical tool, and the ethical question as closure. No existing pedagogical tradition combines all four. The combination is the method.

3.5 The Teacher's Prerequisites

The Know Thyself method cannot be adopted as a technique without being understood as a philosophy. This is its most significant limitation and its most significant demand. A teacher who runs the expected value experiment without understanding Pascal's ontological revolution will present the historical story as decoration - interesting context, easily omitted

when time is short. A teacher who does not understand Mäki's account of isolation will present the ontological question as a warm-up exercise rather than as the method's entry point. A teacher who has never experienced the Smith experiment will describe the institutional turn in experimental economics rather than embodying it.

The method requires the teacher to have read Hacking (1975), Peters (2019), and Robinson (1962) - not as optional enrichment but as prerequisites. It requires familiarity with the experimental economics tradition from Chamberlin through Smith through Gode and Sunder (Guala, 2005). It requires comfort with the philosophy of science literature on models and their limits. And it requires, most importantly, the willingness to have been wrong - to have run an experiment whose outcome surprised them, and to have let that surprise change how they read the model.

This is a high bar. It is the necessary bar. A method that can be adopted without understanding its rationale is a technique. The Know Thyself method is not a technique. It is a way of knowing - about economics, and about oneself. *The method does not change what economics teaches - it changes the order in which the student encounters the question and the answer, and that order changes everything.*

4. Four Case Studies: The Method in Action

The Know Thyself method is not a proposal. It is a documented practice. The four case studies presented in this section are not illustrations of how the method might work. They are records of how it did work - in different courses, with different student populations, addressing different theoretical problems, each one producing the same recognisable pattern: a student who discovered, in their own data, something about themselves that the standard curriculum had given them no reason to expect.

Each case study corresponds to one of the three types of Know Thyself article identified in the project materials: Type A (the functional absence of a tool the student nominally possesses), Type B (a methodological error built into the model itself), and a combined Type B+C (methodological excavation as designed experience). They are presented in the order in which the theoretical argument of this paper requires them, not the order in which they were produced.

A note on transparency: all four case studies involve ad hoc research designs rather than controlled studies. Some are classroom experiments in a narrow sense; others combine surveys,

simulations, valuation tasks, or replication procedures. The samples are convenience samples of enrolled students. The findings are not generalisable in the statistical sense. They are generalisable in the sense that matters for a methodological manifesto: they document a reproducible pedagogical pattern, with sufficient consistency across contexts to constitute evidence that the method produces what it claims to produce.

4.1 Type B: The Ergodicity Case - Freezing Time

> *Reference: Kopczewski & Potocki (2026). The "Two Worlds, Two Urns" Experiment: A Teacher's Reflection on Ergodicity and Economic Methodology. Nonlinear Dynamics, Psychology, and Life Sciences, 30(1), 113–147.*

The hidden ontology. The standard microeconomic model of choice treats a sequence of decisions as equivalent to a single decision evaluated at its expected value. This equivalence holds only if the process is ergodic - if the time average of outcomes for a single agent equals the ensemble average across many agents at a single moment. For multiplicative processes, which describe most real wealth dynamics, this equivalence does not hold. The model freezes time: it removes the arrow, the accumulation, the path dependence that determine where any particular individual actually ends up.

This assumption is never named in the standard curriculum. It is transmitted as an axiom - a feature of the mathematical framework so deeply embedded that it is no longer visible as a choice. Students who learn expected utility theory without learning ergodicity have been handed a model whose most consequential assumption they cannot see.

The paradox. If students were told that the standard model freezes time, most would find the claim implausible. Time is obviously present in intertemporal choice models. Growth models have time indices. How can a model that explicitly models multiple periods be freezing time?

The answer is in what the model does with time: it averages across it. The representative agent who faces a sequence of multiplicative shocks does not accumulate a specific path of outcomes - they experience the ensemble average of all possible paths. A single agent, living through time, experiencing one path with all its compounding and irreversibility, is not the same as the average of all agents across all paths. The model conflates the two. The conflation is the hidden assumption.

The experiment. The Two Urns experiment confronts students with this conflation directly, without naming it first. Two urns are presented: Urn A multiplies the student's stake

by 1.5 with probability 0.5 and by 0.6 with probability 0.5. Urn B offers a certain gain. Students are asked which urn they prefer for a single draw, and which they prefer for a long sequence of draws.

The critical design feature is that the arithmetic expected value of Urn A is positive – the standard model predicts it should be preferred for any number of draws by an expected-value-maximising agent. The geometric mean of Urn A is negative - the ergodic argument predicts that any agent living through a sequence of multiplicative draws from Urn A will, with probability approaching 1, be ruined in the long run. The two predictions point in opposite directions.

Students who have been taught expected value theory consistently prefer Urn A for a single draw and for a sequence of draws - confirming the standard prediction and revealing the hidden ergodic assumption. When the report shows the group distribution and the geometric mean calculation, the response is not confusion but recognition: the model was making a claim about time that was never stated, and that claim is wrong for this situation.

The non-obvious implication. The standard interpretation of behavioural economics findings on risk aversion is that people are irrationally loss-averse - that they deviate from the expected value prediction because of psychological biases that a rational agent would not have. The ergodicity analysis inverts this interpretation: people who are cautious in multiplicative-risk environments are not irrational. They are responding correctly to a world in which path dependence and compounding create asymmetries that the expected-value framework, by design, cannot capture.

The pedagogical implication is not merely that a new concept should be added to the curriculum. It is that a standard interpretation of a large body of empirical findings – the interpretation that frames much of behavioural economics - requires revision. Students who understand ergodicity do not need to be told that the standard model has limits. They can see, in their own data, exactly where the limit is and why.

Limits. The Two Urns experiment is not conducted under strict research procedures. The sample is a convenience sample of enrolled students. The finding that students prefer the arithmetically positive but geometrically negative urn is consistent with the ergodicity hypothesis, but consistent with other explanations as well. The paper states this explicitly, in the text, not in a footnote: *this is a classroom demonstration, not a controlled experiment.* Its function is epistemic provocation, not statistical inference.

4.2 Type A: The Expected Value Case - The Tool That Was Not Used

> *Reference: Kopczewski & Potocki (in progress). You Neglect Probability If You Do Not Know How to Use Expected Value.*

The hidden ontology. The expected value formula rests on an ontological permission that took two centuries to establish: the permission to calculate. As argued in Section 2.2, this permission was not available before Pascal's wager (~1660). The formula without the permission is a dead tool - a symbol that can be reproduced on an exam without activating the habit of mind that makes it useful.

Students who have been taught expected value theory know the formula. They do not, in the functional sense, possess it. The distinction between declarative knowledge (knowing that the formula exists and how to apply it) and procedural knowledge (reaching for it automatically when a decision under uncertainty arises) is the core diagnostic claim of this case study.

The paradox. Probability neglect - the tendency to make decisions under uncertainty as if probabilities were irrelevant - is typically explained in the behavioural economics literature as a consequence of emotional processing: high-affect scenarios (involving health, death, or large losses) trigger affective responses that override analytical reasoning, producing decisions that ignore probability information. The standard remediation is cognitive: make the probability more salient, reduce emotional loading, encourage System 2 thinking.

This case study proposes a different explanation. The effect is not primarily emotional. It is functional: students do not calculate because they do not have the habit of calculating, regardless of the emotional content of the scenario. The relevant mechanism is not affected. It is the absence of a procedural skill that was never actually taught - only described.

The experiment. Five willingness-to-pay scenarios are presented to students, varying systematically in emotional loading: from neutral financial lotteries to scenarios involving health risks, legal outcomes, and mortality. For each scenario, students state their willingness to pay for a change in probability and indicate whether they used intuition or calculation. Every response is timestamped.

The design includes two groups receiving different probability values for the same scenarios, allowing detection of probability sensitivity without requiring explicit calculation checks. A benchmark task - calculating the expected value of one of the lottery scenarios

explicitly - is included at the end, providing a direct measure of whether the student possesses the procedural skill at all.

The finding. The median response time across probability-weighted scenarios is approximately sixty seconds. The minimum time required to perform a rudimentary expected value calculation for any of the scenarios is two to three minutes. The data establish, with a precision that self-report cannot provide, that the students did not calculate. They reached a number by some other means - intuitive estimation, anchoring, affect - and reported it as their willingness to pay.

Critically, the pattern holds across the emotional gradient. The high-affect scenarios (health, mortality) show slightly larger deviations from expected value predictions, consistent with the affect hypothesis. But the neutral financial scenarios show the same pattern of non-calculation. The effect is not explained by affect alone. It is explained by the functional absence of a procedural skill.

The declaration-behaviour gap confirms the diagnosis. When asked after the experiment whether they had calculated, the majority of students in the high-response-time group said yes. The clock said otherwise. This is not deception. It is the characteristic signature of declarative knowledge mistaken for procedural competence: the student believes they calculated because they know how to calculate, not because they did.

The non-obvious implication. The standard policy response to probability neglect is to make probability information more visible - clearer displays, better framing, more salient presentation. This response addresses the symptom without the diagnosis. If the underlying problem is the functional absence of expected value as a habit of mind, making probabilities more visible does not help a student who does not automatically reach for the formula when they see it.

The pedagogical implication is different: the expected value formula must be taught not as a computational tool but as a habit of mind - and that habit requires understanding why the formula exists, what ontological permission it embodies, and what it means to be the kind of agent who reaches for it. The history is not an optional decoration. It is the mechanism by which declarative knowledge becomes procedural competence.

Scope and limits. The claim that the effect is functional rather than emotional rests on the absence of a significant interaction between emotional loading and response time - a null result that requires careful interpretation. The study cannot rule out that emotional processing

is operating in parallel with the functional deficit, producing the same observable outcome by a different route. The paper states this explicitly.

4.3 Type B+C: The Market Equilibrium Case - Nine Stops

> *Reference: Kopczewski & Lisicki (2026). From the Invisible Hand to Digital Twins: Teaching Market Equilibrium as a Journey Through Economic Methodology. (in progress), University of Warsaw.*

The hidden ontology. Alfred Marshall's supply-demand diagram buries its conditions of validity in the *ceteris paribus* clause. Among those buried conditions is the assumption that a specific institutional mechanism - one that aggregates dispersed private information into a public price - exists and operates without friction. The diagram does not name this mechanism. It assumes it. Students who learn the diagram without learning the institution have acquired a belief - that markets clear naturally - without acquiring the understanding of under what conditions, and why, and what happens when the conditions are not met.

The design. The experiment runs on the LabSee.com platform with nine participants - five buyers and four sellers - who first elicit their own willingness-to-pay and willingness-to-accept valuations for a ceramic mug, then trade under four successive treatments: a fundamentals survey, a continuous double auction with human traders only, a continuous double auction pairing each human with a zero-intelligence-constrained (ZI-C) algorithmic twin sharing their exact valuations, and a double auction pairing each human with a naive-learning twin.

The theoretical equilibrium computed from the elicited valuations is $p^* = 25$ PLN, $q^* = 10$ units. The observed transaction prices across four human trading periods are 30, 25, 30, and 24.5 PLN. Convergence is visible and rapid. The unicorn appeared.

The nine-stop debriefing. The experiment is the pretext. The debriefing is the text. A flexdashboard - generated automatically from the session's data - guides students through nine historical and methodological stops, each using their own live results to make visible one assumption that the Marshallian cross conceals.

Stop 1 establishes Smith's intuition of self-organisation - and its incompleteness: Smith did not specify the institution. Stop 2 shows Marshall's formalisation and its cost: the smooth curves of the textbook are an aesthetic choice, not a logical necessity, as the jagged step functions from the students' own valuations make plain. Stop 3 poses McCloskey's unicorn question: has anyone in this room actually observed competitive equilibrium? Stop 4 replicates

Chamberlin's failure using bootstrap simulation on the students' own data. Stop 5 shows Vernon Smith's institutional fix. Stop 6 confronts students with Gode and Sunder's (1993) zero-intelligence result: in period 3, the ZI-C twins outperformed the human traders. Stop 7 reveals the equilibrium as a statistical region rather than a point. Stop 8 gives each student their individual advantage score against their algorithmic twin. Stop 9 separates efficiency from justice through role randomisation.

The digital twin as individual mirror. The most pedagogically consequential element of the design is the individual comparison with the algorithmic twin. Each student can see, using their nickname, exactly how their profit compared with the profit earned by a zero-intelligence algorithm sharing their own valuations. In the session reported in the paper, the aggregate human advantage over ZI-C twins was +106 PLN in period 1, +136 PLN in period 2, −15 PLN in period 3, and +91 PLN in period 4.

The sign reversal in period 3 - the period in which the ZI-C twins briefly outperformed the humans - is the Gode-Sunder result made personal. A student who discovers that a random algorithm with their own preferences beat them in a specific trading period cannot dismiss the finding as an abstract result about other people's markets. It is a fact about their own behaviour, identified by their own nickname, in a session they participated in. The question it forces is not: is the Gode-Sunder result true? The question is: in what periods does my judgment add value, and in what periods does the institution do the work without me?

The non-obvious implication. The standard reading of the experimental economics literature is that it validates market efficiency: given the right institution, prices converge to the competitive equilibrium. This reading is not wrong. It is incomplete. The full reading - available only after nine stops of methodological excavation - is that efficiency and justice are separable problems requiring different institutional solutions, that the equilibrium is a statistical region whose width measures structural uncertainty, and that individual intelligence adds value in precisely the periods when the institution has not yet done its work. These implications cannot be derived from the diagram. They can only be discovered by living through the experiment.

Scope and limits. The session reported in the paper involved nine participants - a sample too small for statistical inference and too specific for generalisation. The ZI-C comparison is suggestive rather than definitive: the sign reversal in period 3 could reflect sampling variation as much as the Gode-Sunder mechanism. The paper states this in the text, not the footnotes, and frames the finding as a demonstration of the methodological claim rather than a test of it.

4.4 Type B: The Altruism Case - The Norm That Beats the Calculation

Reference: Kopczewski & Okhrimenko (2024). The Paradox of Effective Altruism. Journal of Institutional Economics, 20, e41, 1–18.

The hidden ontology. The selfish axiom identified in Section 2.5 has a precise game-theoretic expression: in a one-shot interaction with an anonymous partner, the rational player gives nothing. But this prediction depends on a specific framing - the strategic one. Andreoni and Miller (2002) proposed a different framing: treat payment to self and payment to other as two consumption goods, subject to a budget constraint. Under this framing, the question of rationality changes entirely. It is no longer whether you give zero, but whether your choices are consistent across varying prices of generosity. The test is GARP. And under GARP, altruism is not irrational. The model never prohibited generosity. The curriculum simply never gave it a seat.

The first paradox: rationality is not egoism. The classroom replication of *Giving According to GARP* gives students eight rounds of an allocation problem - tokens to distribute between themselves and an anonymous partner under varying hold and pass values. Students do not know they are in an altruism study. They are solving consumer choice problems under budget constraints, which is exactly what they are doing.

The report shows each student their allocation pattern, their GARP consistency score, and the group distribution across four empirically separable types. Zero-intelligence allocators give randomly - no structure, no norm. Game-theoretic egoists give nothing - the standard prediction, fully consistent with GARP. Deontological altruists give a fixed share regardless of the pass/hold ratio - following a rule, possibly violating SARP, but expressing a stable and recognisable social norm. Effective altruists give everything when the pass value exceeds the hold value, nothing otherwise - maximising aggregate welfare, consistent with both WARP and SARP.

The first finding surprises students trained to expect chaos: the large majority of choices, across all giving types, are GARP-consistent. Altruistic preferences have structure. They are coherent across budget constraints. Being generous does not make you irrational. The selfish axiom was never a logical implication of the model - it was a pedagogical habit.

The second paradox: the atomic calculation loses. The simulation runs the four types through monomorphic and polymorphic societies. The monomorphic result is what the curriculum predicts: a society of welfare-maximising effective altruists accumulates the

highest aggregate wealth - but only under equal initial endowments, the one condition that never obtains in practice. Under unequal endowments, the efficiency criterion produces growing disparity.

The polymorphic result is the one the curriculum cannot anticipate. In mixed societies - where deontological altruists, effective altruists, and zero-intelligence allocators interact - effective altruists are consistently outperformed, individually and collectively, by deontological altruists. The welfare-maximising strategy is exploitable: an agent who gives everything whenever giving is efficient is predictable in ways that make them vulnerable in a world of heterogeneous types and unequal endowments. The deontological rule - give a fixed share, unconditionally, without calculating - is harder to exploit and more robust. The norm that may violate SARP turns out to be the strategy that mixed societies select for. Effective altruism is optimal in a world of effective altruists. In the world that actually exists, it is dominated.

The non-obvious implication. The paper is the first to express deontological and consequentialist ethics as giving strategies with testable emergent properties. The result has a precise pedagogical target: the normative apparatus that economics transmits - maximise aggregate welfare, allocate to highest marginal benefit - produces the best outcome only under conditions that never obtain. The deontological norm that the atomistic framework of Section 2.4 excludes by construction - giving as a social obligation, not a calculation - is not an anomaly to be explained away. It is the strategy that emerges from ecological rationality: not from individual optimisation but from the norms that social interaction evolves when individual calculation is insufficient.

Limits. The classroom experiment uses non-incentivised choices and convenience samples. The simulation results follow from the model's structural assumptions, not from the classroom data directly. The conscientious statement: the experiment shows that altruistic preferences have type-structure legible in individual data. The simulation shows what follows when those types interact at scale. The gap between the individual pattern and the social outcome is where the argument lives. The model excluded the deontological norm not because it fails - but because it was not derived from a calculation.

4.5 The Pattern Across Four Case Studies

Four case studies in different subject areas, with different student populations, addressing different theoretical problems, produce the same recognisable structure:

A hidden assumption that the standard curriculum transmits without naming.

An experiment (or a form of ad hoc study) that makes the assumption visible through the student's own behaviour, before the theory is introduced.

A mirror in which the student sees, with their own nickname, where they stood relative to the group and relative to the model's prediction.

A non-obvious implication that changes not just what the student knows about the model but what questions the student now knows how to ask about it.

A limit stated in the text, not in a footnote, that defines what the case study can and cannot claim.

This structure is not accidental. It is the operationalisation of the five pillars in three different empirical settings. The reproducibility of the structure across settings is the method's primary evidence that it is a method and not a collection of unrelated classroom activities.

The case study is not the proof. The pattern across case studies is the proof.

5. Replication as Pedagogy: The Laboratory Principle

> *"If you cannot replicate it, you do not understand it."*
> - Working maxim, Agent-Based Models course, Jagiellonian University, 2025/26

Reading a scientific article is not the same as understanding it. Most students - and, in practice, many researchers - read for conclusions. They absorb the abstract, scan the figures, note the main findings, and form a judgement about whether the result is plausible. What they do not do, and what the standard curriculum does not ask them to do, is check.

Checking means asking: can I reproduce this? Can I build the model the article describes and obtain what it reports? Can I test whether the result holds when I vary the assumptions the author held fixed? Can I find the operating window - the range of conditions under which the result is true - that the article does not report because the author never tested it?

These are not exotic questions. They are the questions that define scientific competence. They are also, in the standard curriculum, almost never asked - because the tools to ask them were not accessible, the time to ask them was not allocated, and the skills to ask them were not taught.

But there is a prior question, one the standard curriculum almost never asks, because asking it requires seeing the model as something other than a technical instrument. Every model

is a story. Behind the equations lies a compressed account of what human beings are - what they optimise, what they ignore, what they are assumed to be indifferent to. The supply-and-demand model is a story about voluntary exchange and price as sufficient signal. The expected value formula is a story about a calculating agent who treats all uncertainty as quantifiable risk. The ergodic growth model is a story about time - that the future of the average is the average of futures. These stories were not obvious. Decoding them - understanding what kind of human a model presupposes - required years of philosophical and methodological training that the standard curriculum does not provide and does not pretend to provide.

When a student asks can I reproduce this result? they are already asking what assumptions hold this result in place? - and that question is inseparable from what assumptions about me are built into this model? Replication is not merely a technical exercise. It is the first moment at which the student can stand outside the model and read it as a text - a text about human nature, about what is fixed and what is free, about which version of themselves the model recognises and which it renders invisible.

For most of the history of economics education, this reading was accessible only to those who had already spent years inside the methodological tradition. The story behind the model was locked. This has changed. A student who can articulate a precise question - what does this model assume about how I make decisions under uncertainty? - can now, with the right prompting structure and the assistance of a large language model, begin to excavate that assumption on the first encounter. The technical barrier to decoding the story has collapsed. What remains is the only barrier that cannot be automated: scientific curiosity. The student who does not want to know who they are in the model will not find out. The student who does has, for the first time, the tools to look.

The replication laboratory described in this section changes all three conditions simultaneously. The replication protocol is not presented here as an abstract teaching idea. It already exists as a structured prompting artefact: a reusable template that forces the student to separate what is explicit, inferred, assumed, and genuinely ambiguous before any code is written. The prompt is not a technical convenience. It is a methodological document that encodes, in operational form, the epistemological discipline the Know Thyself method requires. Table 1 presents its condensed structure. The full prompt is in Appendix A.

5.1 The Prompt as Methodological Artefact

The prompt system described in this section was developed iteratively over several courses and represents a consolidation of practical experience with AI-assisted laboratory construction. Its central design decision - that no code is written until the ambiguity list from Stage 1 is complete - was not obvious at the outset. It became clear through repeated observation of what happened when students skipped Stage 1: they generated technically correct code that silently resolved the article's underdeterminations in whatever direction the AI's training data suggested, producing implementations that looked rigorous and were epistemologically empty.

The ambiguity list is the epistemological payload of the entire exercise. It is not a list of things the student does not understand. It is a list of things the article does not specify - places where the model's description is genuinely underdetermined, where a choice must be made by any implementor, and where different reasonable choices produce different results. Two programmers implementing the same article could write different code and both be right. The ambiguity list identifies exactly where this is true.

This is a different kind of critical reading than students are typically trained to perform. Standard critical reading identifies logical inconsistencies, unsupported claims, and methodological weaknesses. The ambiguity list identifies something more fundamental: the gaps in specification that peer review did not catch because they are not inconsistencies - they are silences. The article is not wrong. It is incomplete in a specific, consequential way that only becomes visible when you try to build what it describes.

The four-stage sequence enforces a discipline that mirrors the Know Thyself session structure: encounter the article's silences before the code is generated, not after. Stage 1 (extraction) is the experience. Stage 2 (architecture) is the mirror - the moment at which the student must decide what questions the laboratory will allow them to ask. Stage 3 (code) is the theory rendered operational. Stage 4 (extension guide) is the ethical question in a different register: having built this laboratory, what would you need to know next?

Table 1. Replication laboratory protocol - condensed structure

Component	Operational specification
Philosophy of the laboratory	Three questions every laboratory must allow the student to answer: What is certain?: What does the article actually prove - not merely suggest? What is assumed?: Where did the implementor choose, because the article was silent?

Component	Operational specification
	What is fragile?: Under which changes in assumptions do the results disappear or reverse? Rewarded behaviour: scepticism. The dashboard exists to make the article fail, not to display it.
Four-stage protocol	Stage 1 - Extraction: Classify every element of the article: (a) explicit, (b) inferred, (c) assumed by the implementor, (d) genuinely ambiguous. Produce the ambiguity list. No code until this stage is complete. Stage 2 - Architecture: Design the tab structure. For each tab, one sentence: what will the student learn and what research question can they ask? Stage 3 - Full code: Write the complete .Rmd file. Every comment explains WHY this step matters for the model, not only what the code does. Stage 4 - Extension guide: Document how new experiments can be added. This is a research agenda, not a maintenance note.
Three article modes	Module A - Simulation: Article describes a computational model. No empirical data. Key tools: parameter sweep, seed sensitivity, phase boundary identification. Module B - Empirical with data: Article has published data and/or code. Key tools: specification sweep, coefficient plot, subgroup heterogeneity, outlier sensitivity. Module C - Reverse engineering: Article has only tables of results. No data, no code. Key tools: Approximate Bayesian Computation (ABC), copula modelling, power analysis, uncertainty map.
Evidence labelling	[ART]: Value taken directly from the article (table, figure, text). [IMP]: Assumption made by the implementor where the article is silent. [SYN]: Value generated by ABC or other reconstruction method. Every number in the dashboard carries one of these three labels. Unlabelled numbers are not permitted.
Uncertainty map	Required in every dashboard. A visualisation of the confidence level for each key claim in the article: 0.0 - article is silent on this claim 0.5 - claim is consistent with published results but not uniquely determined by them 1.0 - claim is stated explicitly and directly supported by reported data The map forces the student to evaluate not whether claims are true, but how well-supported they are by what the article provides.

The full replication prompt - including R/Shiny implementation conventions, sidebar design, HTML notation rules, and the complete checklist - is provided in Appendix A and in the online repository.

5.2 The Replication Crisis as Pedagogical Opportunity

The replication crisis in the social sciences has produced an extensive literature on why published findings fail to survive independent verification: publication bias, underpowered studies (Button et al., 2013), researcher degrees of freedom, and the file-drawer problem (Open Science Collaboration, 2015; Ioannidis, 2005; Camerer et al., 2016). This literature is primarily methodological - it diagnoses the problem and proposes remedies at the level of research practice.

The Know Thyself method takes a different angle: the replication crisis is not only a problem for researchers. It is a pedagogical opportunity. A student who tries to replicate a published model and discovers that it cannot be fully replicated - not because the model is wrong, but because the article is underdetermined - has learned something about scientific knowledge that no lecture on research methodology can convey. They have experienced the underdetermination. They have not been told about it.

This is the description-experience gap (Hertwig & Erev, 2004) applied to scientific reading. An article describing a model communicates its results. What it does not communicate - what cannot be communicated by description alone - is the experience of building the model: the decisions made at every point where the article was silent, and the ways in which those silent decisions change the results. The gap between reading a simulation article and building the model it describes is the same structural gap as the gap between being told the expected value formula and discovering, in one's own timestamped data, that one did not use it.

5.3 Three Levels of Replication

Not all articles present the same replication challenge. The appropriate tools depend on what the article provides. The framework distinguishes three modules within a common protocol structure. The common structure ensures that the student's inquiry follows the same epistemological sequence regardless of what the article provides: extract what is known, identify what is assumed, test what is claimed, map what is uncertain.

Module A - Simulation articles.

When an article describes a computational model - agent-based, cellular automata, system dynamics - with no empirical data, the replication challenge is implementation fidelity. The model's rules may appear determinate. They rarely are. The ambiguity list typically identifies at minimum: boundary conditions, step definition, initialisation procedures, and

convergence criteria. Each of these silences has measurable consequences. The student who discovers, through building the model, that their implementation produces a power law exponent that differs from the article's reported value by 0.3 - depending on the boundary handling - has not found an error. They have found a scope condition: a specification of where the article's result holds and where it does not. The article did not provide this. The laboratory produced it.

Module B - Empirical articles with published data and code.

When an article has published its data and code, the replication challenge shifts from implementation fidelity to specification robustness. Every empirical analysis involves choices that appear obvious or conventional: which control variables to include, how to handle outliers, which functional form to specify. These choices are not arbitrary - experienced researchers make them on principled grounds. But they are choices, and different principled choices produce different results. The Module B laboratory implements a specification sweep: a systematic variation of the choices the author made, producing a coefficient plot that shows the estimated effect across a range of defensible specifications. A result that holds across all specifications is robust. A result that disappears when one control variable is added is fragile. The student who has produced this plot has understood the result's stability in a way that the article's tables do not communicate.

Module C - Reverse engineering from published statistics.

The most methodologically innovative module addresses the most common case in the pre-open-science literature: the article provides only tables of results, methodological descriptions, and figures. No data. No code. The question is what a careful reader can do with this - and the answer is more than is usually assumed.

The published statistics in an article's tables - means, standard deviations, regression coefficients, standard errors, sample sizes, p-values - are not merely summaries of data. They are constraints on the data. Any dataset that produced those statistics must fall within a specific region of the space of possible datasets. Approximate Bayesian Computation (ABC; Beaumont et al., 2002; Toni et al., 2009) provides the operational framework for characterising this region: the procedure generates many candidate datasets from the prior distribution, computes summary statistics for each, and retains the candidates whose statistics fall within a specified tolerance of the article's published values. The retained candidates constitute the approximate posterior - the distribution of datasets consistent with the article's claims.

The reconstruction reveals what the article does not report: the width of the posterior (how many very different datasets are equally consistent with the published results) and the fragility of the conclusions (which claims hold across the entire posterior and which hold only in a subset of it). A narrow posterior means the results are highly informative: few conditions could have produced them. A wide posterior means the results are weakly informative: many very different conditions are equally consistent with what was published. Neither finding is a criticism of the article. Both are facts about what the article's design allows one to conclude - facts the article does not contain and that only the laboratory can produce.

Power analysis completes the Module C toolkit. For any reported result, the laboratory computes the minimum detectable effect at conventional power levels given the sample size. For null results - findings of no significant effect - it computes how large an effect could have gone undetected. A student who has produced a power curve for an article's main null finding and has seen that the study had 40% power to detect the claimed null has understood something the article's conclusions section does not say: the null result is uninformative. It is consistent with the effect being real and the study being underpowered. The laboratory makes this visible. The article did not.

5.4 The Uncertainty Map as Epistemic Instrument

Every replication laboratory produced under this protocol includes, as a required element, an uncertainty map: a visualisation of the confidence level for each of the article's key claims. The confidence scale runs from zero - the article is silent on this claim; it rests entirely on the implementor's assumptions - to one - the claim is stated explicitly and directly supported by the reported data. Claims in between carry intermediate confidence, with the specific level depending on whether the claim is inferred from context, reconstructed from statistics, or supported by results that held across the specification sweep.

The uncertainty map forces the student to evaluate, for each of the article's claims, not whether the claim is true but how well-supported it is by what the article provides. This is a different cognitive act from reading critically. Critical reading identifies what the article says and asks whether the argument is valid. The uncertainty map asks: given what the article says, what can I actually know from it? The distinction is between evaluating an argument and evaluating its evidential base - and it is a distinction the standard curriculum rarely makes operational.

Students who have produced uncertainty maps for several articles develop, through the exercise, a calibrated sense of what different confidence levels mean in practice. They learn that a claim with confidence 0.3 is not necessarily false - it may be well-motivated theoretically - but that it is not established by the article that asserts it. They learn that a claim with confidence 1.0 is established by that article's data, within that article's sample, under that article's specifications - and may still be fragile in the Module B sense. Calibration of this kind is not a skill that reading produces. It is a skill that building produces.

5.5 The Democratisation of Replication

Before AI-assisted replication, the practical barrier to checking a published result was prohibitive for most students and many researchers: obtain access to the data, master the software, understand the code, reproduce the analysis, test the robustness. The prompt system described here does not eliminate these steps. It makes them accessible to any student who can read an article carefully enough to produce an ambiguity list.

This is a significant claim about what AI changes in scientific education. The change is not that AI writes the code - though it does, and this matters. The change is that a well-structured prompt converts the epistemological discipline of replication into a reproducible, transferable procedure. The prompt encodes, in operational form, what it means to read a scientific article as a scientist rather than as a student: to ask what is certain, what is assumed, and what is fragile, and to build the tool that answers those questions.

Any student, in any course, with access to any scientific article and to an AI assistant, can now produce a replication laboratory. The barrier is not technical. It is epistemological - and it is precisely the barrier that Stage 1 of the protocol addresses. The student who cannot produce the ambiguity list has not read the article in the sense that replication requires. The student who can produce it has understood what the article does and does not establish - and has the tool to find out.

This is the connection to the broader argument of this paper. The Know Thyself method, in the classroom, produces students who see themselves in data before they receive the theory. The replication laboratory, applied to the literature, produces students who see the article's silences before they accept its conclusions. Both procedures activate the same cognitive mechanism: the experience of incompleteness that Golman and Loewenstein (2018) identify as the trigger of genuine curiosity. Both produce the same kind of knowledge: not information about a result, but a changed relationship to the question the result was supposed to answer.

A well-structured prompt, applied to any scientific article, is sufficient to turn that article into a laboratory. What it cannot replace is the student who reads carefully enough to know what questions to ask - and that student is what the Know Thyself method is designed to produce.

6. The Socratic Tool: Artificial Intelligence as Infrastructure, Mirror, and Collective Memory

6.1 AI as the Removal of Infrastructural Friction

Artificial intelligence did not create the Know Thyself method. The method existed before AI entered the classroom. Its logic - experience before theory, mirror in data before concept, question before answer - was already present in the experiments, reports, and teaching practices developed over many years. What AI changed was not the epistemology of the method, but the resistance of the medium.

For a long time, the main obstacle was not the absence of ideas, data, or pedagogical intention. It was the material difficulty of turning those intentions into working infrastructure. Each experiment required code. Each report required formatting. Each dashboard required debugging. Each replication exercise required technical scaffolding that often consumed more time than the methodological idea itself. The teacher had to be not only a teacher and an economist, but also a programmer, designer, data cleaner, and technical support system. The method was possible, but it demanded too much heroic labour.

AI changes this condition. It lowers the cost of transforming an epistemic provocation into a working educational object: a survey, a simulation, a personalised report, a flexdashboard, a replication protocol, or a critical reading laboratory. It does not replace the teacher's judgement. It does not decide what the experiment should mean. It does not identify the hidden ontology of the model. But it makes it much easier to give that judgement an operational form.

In this sense, AI completes the method only in a practical, infrastructural sense. It does not give the method its logic. It gives it its machinery. It allows a pedagogical idea that previously existed in fragments - classroom experiments, improvised reports, hand-built dashboards, partial replications - to become reproducible, scalable, and shareable.

This distinction matters. If AI is treated as a source of answers, it works against the Know Thyself method. It gives students polished responses before they have experienced the question.

It restores the very sequence the method tries to break: answer before experience, explanation before doubt, fluency before understanding. But if AI is treated as infrastructure, its role is different. It helps construct the mirror. It helps expose assumptions. It helps transform a scientific article into a replication laboratory. It helps students see not only what a model says, but what must be supplied, assumed, cleaned, coded, or interpreted before the model can speak.

The deepest implication is therefore not technological but pedagogical. AI makes the Know Thyself method easier to implement at the same moment when the method becomes more necessary. In an AI-saturated classroom, students can produce fluent answers without ever encountering the resistance of the problem. The task of teaching is therefore not to forbid AI, nor to celebrate it uncritically, but to place it in the correct sequence. First the question. First the ambiguity. First the student's own decision, error, hesitation, or assumption. Only then the tool.

AI did not invent Know Thyself. It removed part of the friction that had kept the method from appearing in its full operational form. The danger is that the same technology can also remove the friction from thinking itself. The method's role is to prevent that. Its purpose is not to make learning smoother, but to make the right difficulty visible.

Once this infrastructural role is clarified, three layers of AI use can be distinguished. They are not variations on a single theme. They are structurally different claims, carrying different degrees of certainty: AI as machinery for building the method, AI as a Socratic partner in replication, and AI as a speculative problem of collective memory and epistemic diversity. We take them in that order, from the most certain to the most speculative.

6.2 AI as Coder - Documented, Operational

The Know Thyself method is, among other things, a methodology of mirrors: students must see themselves in data before theory becomes relevant. AI does not change this logic. It makes the mirrors easier to build: personalised, interactive, generated after the task rather than before it, and available to instructors who are not programmers.

Generative AI has lowered that barrier substantially.

The prompt system described in Section 5 and provided in full in Appendix A demonstrates this concretely. Two distinct prompt architectures are now operational: one for building the Know Thyself session report - the personalised mirror that students see after each experiment - and one for building replication laboratories from scientific articles. Both follow

the same structural logic: the prompt encodes the pedagogical philosophy, not only the technical requirements. It specifies what questions the dashboard should raise rather than answer, where the session should end (open questions for discussion, not a summary of results), and how the student's individual data should be positioned relative to the group aggregate.

The act of writing the prompt has become a pedagogical act in its own right. A teacher who writes the prompt for a Know Thyself session must decide, in operational terms, what the student should encounter first, what should be withheld until the data are visible, and what question should be left unanswered at the end. These are not technical decisions. They are decisions about the epistemological sequence of a lesson - and they force the teacher to have understood the method at the level of its logic, not only its procedure.

This is documented practice. The prompts are published. Any teacher can adapt them. The claim here is narrow, verifiable, and verified: AI-assisted dashboard construction scales the Know Thyself method to instructors who are not programmers, and the prompt that produces the dashboard is itself a record of the pedagogical decisions the method requires.

6.3 AI as Socratic Partner - Demonstrated, Documented

The second layer is no longer emerging. It is demonstrated in the protocol described in Section 5 - and demonstrated means something specific: the four-stage replication protocol exists, has been used in courses, and produces the epistemic behaviour it is designed to produce. Students who complete Stage 1 before writing code discover things about scientific articles that they did not know before attempting the replication. Students who skip Stage 1 produce implementations that look rigorous and are epistemologically empty. The difference is observable and consistent.

What AI makes possible in this layer is not replication per se - replication has always been possible for researchers with sufficient time and expertise. What AI makes possible is replication as a standard pedagogical tool: accessible to any student, applicable to any article, completing in a single session rather than a semester. The constraint that makes this possible is also the constraint that makes it Socratic: no code before the ambiguity list.

The Socratic parallel is precise. Socrates did not give answers. He asked questions until his interlocutor discovered what they did not know. The replication protocol does the same, at scale: the four-stage sequence is a structured interrogation of the article, and AI is the instrument that makes the interrogation productive rather than merely time-consuming. But

- and this is the constraint that gives the method its bite - AI can only play this role if the student has already done Stage 1. A student who asks AI to *write the model* without having produced the ambiguity list has not interrogated the article. They have asked a courtier to describe the clothes. The ambiguity list is the condition under which the child can say what they see.

The operative rule - ununderstood code is not your code - is not a technical standard. It is an epistemological one. If AI has written something the student cannot explain, the student asks until they can explain it, or rewrites it until it is theirs. This rule applies to every modelling choice in the implementation, including the choices AI made silently in the places where the article was silent. Those silent choices are the article's hidden assumptions made operational - and making them visible is the point of the entire exercise.

Three things that the replication protocol produces, which reading alone cannot: the ambiguity list (the article's silences made explicit), the uncertainty map (the article's claims calibrated by their evidential support), and the fragility finding (the result's operating window, which the article does not report because the author never tested it). All three are outputs of the student's labour, not the AI's. AI wrote the code. The student produced the knowledge.

6.4 AI as Collective Memory - A Speculative Boundary

The third layer is where the argument becomes most ambitious, and where we must be most careful about the boundary between what the literature supports and what we are proposing.

Scott Page's Diversity Prediction Theorem states a mathematical identity: the mean squared error of a crowd's aggregate prediction equals the average individual error *minus* the diversity of the crowd's predictions (Page, 2007). This is not a metaphor. It is an algebraic fact. The implication is precise: a perfectly homogeneous crowd - one in which every member makes the same prediction - achieves exactly the accuracy of any one of its members, however wrong that may be. Diversity is not a social virtue decorating the aggregate; it is a structural component of its accuracy.

James Surowiecki's argument in The Wisdom of Crowds (2004) operationalises this insight: crowds are wise when their members hold genuinely independent, diverse judgements. They fail - they become mobs rather than oracles - when social influence, informational cascades, or conformity pressure homogenise individual assessments before aggregation. The crowd's intelligence is a function of the independence and diversity of its inputs.

Large language models are, among other things, interfaces to the accumulated textual output of human thought. They are trained on corpora that represent, however imperfectly, the written knowledge of a civilisation. In Page's terms, they approximate an extraordinarily large aggregate. The educational question is not whether AI is intelligent. The question is what happens to learning when students stop contributing independent judgement to that aggregate and begin merely recycling its outputs.

Jacek Dukaj poses this question with a precision that academic prose has not yet matched. In *Perfekcyjna niedoskonałość* (2004) - *Perfect Imperfection* - he imagines post-human intelligences for whom perfect information processing has eliminated the friction of uncertainty. They are no longer wrong in ways that are interesting. The argument is not that they are stupid. It is that they have lost the productive imperfection that generates genuinely new knowledge: the mistake that reveals an assumption, the confusion that signals an unresolved problem, the anomaly that breaks the model.

The student who uses AI as an oracle - who inputs a problem and accepts the output without interrogation - does not merely fail to learn. They reduce the independence of the inputs on which AI-mediated knowledge work depends. They copy from the aggregate back into the aggregate, closing a loop that amplifies whatever biases and gaps the aggregate already contains. In Page's terms, they subtract diversity.

The student who passes through the Know Thyself protocol does something structurally different. They participate in an experiment before they know the theory. They see their own choices - often inconsistent, often surprising to themselves - displayed against a distribution of choices made by people who are different from them. They produce an ambiguity list for an article and find a silence that nobody in the room had noticed before. They discover that their ZI-C twin - an algorithm with their own preferences and a single constraint - outperformed them in the third trading period.

Each of these experiences produces an epistemic event that is, in Page's sense, genuinely diverse: it is specific to that student, in that room, on that day, with those data. It does not reproduce what is already in the aggregate. It adds to it.

This is why the Know Thyself method matters in an age of AI - not because AI is dangerous in itself, but because AI-mediated knowledge work depends on the diversity and independence of the human questions brought to it. The method is, among other things, a mechanism for producing epistemic inputs that the aggregate does not already contain.

The institutional parallel.

Vernon Smith's insight about the continuous double auction applies here at a different level. In Smith's account, the institution does most of the work: it aggregates dispersed private information into a public price without requiring any individual to be particularly rational. The institution is the intelligence. But that institution depends, for its function, on there being genuinely private information to aggregate. A market in which every participant has been told the equilibrium price in advance does not discover the equilibrium - it ratifies it. The institution's intelligence is a function of the diversity of its inputs.

AI can be understood as an institution in Smith's sense: it aggregates. But aggregation of homogeneous inputs is noise amplification, not intelligence. The Know Thyself student – the one who has experienced the question before receiving the answer, found the ambiguity an article concealed, and seen themselves in data that surprised them - is, in this framework, a source of signal. Not because they are smarter than the aggregate. Because they have done something the aggregate has not done: they have been wrong about something specific, in a specific situation, and they have seen it.

The exoskeleton argument.

AI as an intellectual exoskeleton - a tool that amplifies rather than replaces human capacity - is a useful metaphor, but it requires a precise condition to be non-trivial: the human inside the exoskeleton must be doing something the exoskeleton cannot do alone. An exoskeleton attached to a mannequin does not walk.

The human contribution to the AI-assisted educational process is not merely providing prompts. It is providing the epistemic diversity - the irreducibly personal experience of having been wrong about something specific, in a specific situation, with specific data - that keeps the aggregate from converging to its own reflection. The prompt in Appendix A can be used by anyone. The ambiguity list it requires cannot be produced by anyone who has not read the article carefully enough to find the silences. The uncertainty maps it mandates cannot be produced by anyone who has not understood what the article claims and what it merely implies. These are human acts. They are also, in the current state of AI, acts that AI cannot perform on the student's behalf - not because AI lacks the capability, but because performing them on the student's behalf defeats the purpose.

6.5 The Boundary

The layers discussed in this section carry different epistemic statuses, and we have stated this from the outset. The coding layer is documented and verifiable: the prompts exist, they work, they are published. The Socratic-replication layer is demonstrated: the protocol exists, has been used, and produces the behaviour it is designed to produce. The collective-memory layer is a philosophical thesis: it is grounded in Page's theorem, Surowiecki's argument, and Smith's institutional economics, extended by Dukaj's speculative fiction. It is not measured. It cannot currently be measured. We state it as a thesis - explicitly, in the text - because the transparent framing of a speculative claim is more useful to the reader than its concealment behind hedged language.

What we can say, without speculation, is this: a student who has seen themselves in data, who has produced an ambiguity list for an article they replicated, and who has understood why their algorithm-twin beat them in a specific trading period, has a different relationship to artificial intelligence than a student who has not. They are not afraid of it. They are not naive about it. They know, with the precision that only personal experience can deliver, when the institution suffices and when the human is still required. *That knowledge is not in the aggregate. It has to be earned.* AI-mediated knowledge work is only as rich as the diversity of the humans who question it - and the Know Thyself method is, among other things, a machine for producing that diversity.

7. Limits and One Implication

This paper has made several claims. Some are narrow and documented. Some are demonstrated in practice but not formally measured. One is a philosophical thesis. The limits of each have been stated as they arose - in the text, not in footnotes. This section collects them in one place, and then turns to the single practical implication that follows from everything that precedes it.

7.1 Three Limits

The teacher prerequisite is high - and is not negotiable. The Know Thyself method cannot be adopted as a technique without being understood as a philosophy. A teacher who runs the expected value experiment without understanding Pascal's ontological revolution will present the history as decoration. A teacher who has never experienced the Smith experiment will describe the institutional turn in experimental economics rather than embody it. A teacher

who builds the replication laboratory without having read Mäki, Robinson, and Peters will produce dashboards without understanding what they are designed to reveal.

This is not an argument for adding prerequisites to teacher training - though it would not hurt. It is an argument for transparency about what the method requires. The barrier to entry is genuine. It is the barrier of having been wrong about something in one's own field, having seen it in data, and having let it change how one reads a model. That experience cannot be delegated to a prompt or acquired from a reading list. It must be lived. The prompts in Appendix A are tools for teachers who have already crossed that threshold. They are not shortcuts to it.

There is no RCT - and there will not be one. The four case studies in Section 4 document a reproducible pattern across different student populations, different subject areas, and different theoretical problems. They establish that the method produces a recognisable epistemic event - the student's discovery of their own assumption in their own data - with sufficient consistency to constitute evidence. They do not establish, and cannot establish, the size of the effect relative to a control condition, the persistence of the effect over time, or the generalisability of the pattern to populations outside the author's courses.

A randomised controlled trial of the Know Thyself method is not, in practice, possible: the method requires the teacher to have internalised it, and internalisation cannot be randomly assigned. This is not an excuse. It is a structural feature of any pedagogical intervention whose mechanism depends on the teacher's own epistemic state. We state it here because the standard evidentiary criterion for educational interventions - the RCT - is not appropriate for this kind of claim, and pretending otherwise would be the methodological sin the method is designed to prevent. The appropriate evidential standard for a proof of concept with documented reproducibility is: does the pattern hold across independent implementations? That question is answerable, and the answer, across four case studies, is yes.

The replication protocol's scope conditions are not yet mapped. The four-stage replication protocol has been used in courses on agent-based modelling and selected economics courses. It works in those contexts. Whether it scales to introductory courses, larger enrolments, disciplines with different methodological conventions, or students with less quantitative backgrounds remains unknown. Module C - the reverse-engineering approach using Approximate Bayesian Computation - is the least tested of the three modules. It has been implemented and produces the intended epistemic behaviour. Its operating window - the range of articles, student backgrounds, and course structures for which it is appropriate - requires further documentation.

We flag this not to weaken the claim but to define it precisely. A proof of concept that states its scope conditions explicitly is more useful than a generalisation that conceals them. The protocol is real, operational, and published. Its limits are the limits of documented practice, not of theoretical aspiration.

7.2 One Concrete Implication

Everything in this paper - the hidden ontology of economic models, the five pillars of the method, the four case studies, the replication laboratory, and the argument about AI as infrastructure and epistemic diversity - converges on a single pedagogical recommendation. It is not a system redesign. It is not a curriculum reform. It requires no grant, no platform, no programming skills, and no institutional permission.

Change the order. Before you teach the expected value formula - ask your students to make a decision under uncertainty and record how long it takes. Before you draw the supply-demand diagram - ask your students whether they have ever seen prices emerge from the interaction of buyers and sellers, and what happened. Before you introduce the representative agent model - ask your students whether they think the average outcome across a population at one moment tells them anything reliable about what will happen to them personally over time.

These questions do not require a full experiment. They do not require a platform. They require thirty seconds and the willingness to leave the question open until the student has felt it - until the gap between what they assumed and what the data, answer, simulation, or reconstruction will show has had time to register as a question that belongs to them.

The theory can wait. It has been waiting in textbooks for decades. What it cannot do, from inside the textbook, is reach a student who has not yet asked the question it answers. The Know Thyself method bets on a simple mechanism: that a student who has experienced the failure of a tool they thought they had will reach for the correct tool - and understand why they are reaching for it - with a motivation that no amount of lecturing can produce.

This is not a claim about pedagogy. It is a claim about knowledge. The difference between knowing the formula and knowing when to reach for it is not a difference in information. It is a difference in the relationship between the student and the question. That relationship is changed by experience, not by description. And experience, in the Know Thyself method, begins with one move: the teacher who asks the question before giving the answer.

Monsieur Jourdain discovered he had been speaking prose for forty years and was delighted. The economics teacher who discovers they have been assuming ergodicity for forty years faces a harder reckoning. But both discoveries begin the same way: with a question that reveals, gently and irreversibly, that the world contains something the speaker did not know they were doing.

The method does not ask the teacher to have all the answers. It asks the teacher to have the question - and to trust that the student, once they have seen themselves in the data, will want the answer badly enough to learn it properly. *Change the order. Ask the question that hurts. Wait for the student to feel it. Then teach.*

AI statement: This paper was written in conversation with an AI assistant that contributed no ideas and originated no arguments. Its role was what the poet demanded of language itself: that the supple tongue say all the head can think - no more, no less. The author supplied the head. The machine supplied the tongue. Responsibility for both remains entirely with the author.

"I chodzi mi o to, aby język giętki powiedział wszystko, co pomyśli głowa."
"I want the supple tongue to say all that the head can think."

- Juliusz Słowacki, Beniowski, Pieśń V

References

Andersen, H.C., 1837. The emperor's new clothes. In Fairy tales told for children. C. A. Reitzel.

Andreoni, J., Miller, J., 2002. Giving according to GARP: An experimental test of the consistency of preferences for altruism. Econometrica, 70(2), 737–753.

Ariely, D., Loewenstein, G., Prelec, D., 2003. Coherent arbitrariness: Stable demand curves without stable preferences. Quarterly Journal of Economics, 118(1), 73–106.

Backhouse, R.E., 2008. The puzzle of modern economics. Journal of Economic Methodology, 15(1), 1–16.

Beaumont, M.A., Zhang, W., Balding, D.J., 2002. Approximate Bayesian computation in population genetics. Genetics, 162(4), 2025–2035.

Blaug, M., 2003. The formalist revolution of the 1950s. Journal of the History of Economic Thought, 25(2), 145–156.

Button, K.S., Ioannidis, J.P.A., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S.J., Munafò, M.R., 2013. Power failure: Why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience, 14(5), 365–376.

Camerer, C.F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., et al., 2016. Evaluating replicability of laboratory experiments in economics. Science, 351(6280), 1433–1436.

Camerer, C.F., Loewenstein, G., Rabin, M. (Eds.), 2004. Advances in Behavioral Economics. Princeton University Press.

Chamberlin, E.H., 1948. An experimental imperfect market. Journal of Political Economy, 56(2), 95–108.

Colander, D., 2005. The making of an economist redux. Journal of Economic Perspectives, 19(1), 175–198.

Dukaj, J., 2004. Perfekcyjna niedoskonałość [Perfect Imperfection]. Wydawnictwo Literackie.

Earle, J., Moran, C., Ward-Perkins, Z., 2017. The Econocracy: The Perils of Leaving Economics to the Experts. Manchester University Press.

Friedman, M., 1953. The methodology of positive economics. In Essays in Positive Economics (pp. 3–43). University of Chicago Press.

Gibbard, A., Varian, H.R., 1978. Economic models. Journal of Philosophy, 75(11), 664–677.

Gigerenzer, G., 2008. Rationality for Mortals: How People Cope with Uncertainty. Oxford University Press.

Gigerenzer, G., Hertwig, R., Pachur, T. (Eds.), 2011. Heuristics: The Foundations of Adaptive Behavior. Oxford University Press.

Gode, D.K., Sunder, S., 1993. Allocative efficiency of markets with zero-intelligence traders: Market as a partial substitute for individual rationality. Journal of Political Economy, 101(1), 119–137.

Golman, R., Loewenstein, G., 2018. Information gaps: A theory of preferences regarding the presence and absence of information. Decision, 5(3), 143–164. https://doi.org/10.1037/dec0000068

Guala, F., 2005. The Methodology of Experimental Economics. Cambridge University Press.

Hacking, I., 1975. The Emergence of Probability. Cambridge University Press.

Hands, D.W., 2001. Reflection Without Rules: Economic Methodology and Contemporary Science Theory. Cambridge University Press.

Hausman, D.M., 1992. The Inexact and Separate Science of Economics. Cambridge University Press.

Hertwig, R., Erev, I., 2004. The description–experience gap in risky choice. Trends in Cognitive Sciences, 8(12), 517–523.

Ioannidis, J.P.A., 2005. Why most published research findings are false. PLOS Medicine, 2(8), e124.

Kahan, D.M., Landrum, A.R., Carpenter, K., Helft, L., Jamieson, K.H., 2017. Science curiosity and political information processing. Political Psychology, 38(S1), 179–199.

Kahneman, D., Knetsch, J.L., Thaler, R.H., 1990. Experimental tests of the endowment effect and the Coase theorem. Journal of Political Economy, 98(6), 1325–1348.

Kant, I., 1785/1997. Groundwork of the Metaphysics of Morals (M. Gregor, Trans.). Cambridge University Press.

Kolb, D.A., 1984. Experiential Learning: Experience as the Source of Learning and Development. Prentice Hall.

Kopczewski, T., Lisicki, J., (in progress). From the invisible hand to digital twins: Teaching market equilibrium as a journey through economic methodology.

Kopczewski, T., Okhrimenko, I., 2024. The paradox of effective altruism. Journal of Institutional Economics, 20, e41. https://doi.org/10.1017/S1744137424000146

Kopczewski, T., Potocki, T., 2026. The "Two Worlds, Two Urns" Experiment: A Teacher's Reflection on Ergodicity and Economic Methodology. Nonlinear Dynamics, Psychology, and Life Sciences, 30(1), 113–147.

Kopczewski, T., Potocki, T. (in progress). You neglect probability if you do not know how to use expected value.

Mäki, U., 1992. On the method of isolation in economics. Poznan Studies in the Philosophy of the Sciences and the Humanities, 26, 319–354.

Mäki, U., 2009. MISSing the world: Models as isolations and credible surrogate systems. Erkenntnis, 70(1), 29–43.

Marshall, A., 1890. Principles of Economics. Macmillan.

McCloskey, D.N., 1985. The Rhetoric of Economics. University of Wisconsin Press.

Mercier, H., Sperber, D., 2017. The Enigma of Reason. Harvard University Press.

Mirowski, P., 2013. Never Let a Serious Crisis Go to Waste: How Neoliberalism Survived the Financial Meltdown. Verso.

Molière [Jean-Baptiste Poquelin], 1670. Le Bourgeois Gentilhomme. Paris.

Open Science Collaboration, 2015. Estimating the reproducibility of psychological science. Science, 349(6251), aac4716.

Page, S.E., 2007. The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies. Princeton University Press.

Paul, L.A., 2014. Transformative Experience. Oxford University Press.

Peters, O., 2019. The ergodicity problem in economics. Nature Physics, 15(12), 1216–1221.

Plato, 1997. Apology (G. M. A. Grube, Trans.), in: Cooper, J.M., Hutchinson, D.S. (Eds.), Plato: Complete Works (pp. 17–36). Hackett.

Plott, C.R., 1973. Path independence, rationality, and social choice. Econometrica, 41(6), 1075–1091.

Robinson, J., 1962. Economic Philosophy. Aldine.

Rodrik, D., 2015. Economics Rules: The Rights and Wrongs of the Dismal Science. W. W. Norton.

Shiller, R.J., 2017. Narrative economics. American Economic Review, 107(4), 967–1004.

Słowacki, J., 1841. Beniowski. Pieśń V.

Smith, A., 1759. The Theory of Moral Sentiments. A. Millar.

Smith, A., 1776. An Inquiry into the Nature and Causes of the Wealth of Nations. W. Strahan and T. Cadell.

Smith, V.L., 1962. An experimental study of competitive market behavior. Journal of Political Economy, 70(2), 111–137.

Smith, V.L., 2003. Constructivist and ecological rationality in economics. American Economic Review, 93(3), 465–508.

Smith, V.L., 2008. Rationality in Economics: Constructivist and Ecological Forms. Cambridge University Press.

Surowiecki, J., 2004. The Wisdom of Crowds. Doubleday.

Thaler, R.H., Sunstein, C.R., 2008. Nudge: Improving Decisions About Health, Wealth, and Happiness. Yale University Press.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., Stumpf, M.P.H., 2009. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. Journal of the Royal Society Interface, 6(31), 187–202. https://doi.org/10.1098/rsif.2008.0172

Varian, H.R., 2010. Intermediate Microeconomics: A Modern Approach (8th ed.). W. W. Norton.

Annexes

Appendix A. The Know Thyself Prompt Library and Dashboard Repository

This appendix documents the online repository that accompanies the paper. The repository contains the prompt library, implementation conventions, and reusable workflow instructions for constructing the Know Thyself dashboards, replication laboratories, and agent-based model tools discussed in the main text. Repository URL: https://github.com/tomvar/know-thyself-prompt-library

The repository should be read in two ways. First, it is a transparency record: it documents how artificial intelligence was used as a writing, coding, and methodological infrastructure. Second, it is a practical toolkit: it allows readers to reproduce and adapt the workflows by which classroom results, surveys, simulations, scientific articles, and model files are turned into interactive dashboards.

The repository is intentionally a living artefact. The paper refers to a specific version of the prompt library; the GitHub repository preserves its subsequent evolution. This distinction is important. The archived release supports reproducibility. The living repository supports methodological development.

A.1 Relation to the main argument of the paper

The appendix belongs most directly to Sections 5 and 6 of the paper. Section 5 introduces the replication laboratory as a pedagogical extension of the Know Thyself method: scientific articles are not only read but reconstructed, tested, and made fragile on purpose. Section 6 explains the role of AI in this process: AI did not create the method, but lowers the infrastructural cost of building surveys, dashboards, simulations, and replication laboratories.

The prompt library operationalises these claims. It turns the paper's epistemological sequence - experience before theory, mirror in data before concept, question before answer - into reusable instructions. In the ad hoc research workflow, participants first generate data through a survey, experiment, simulation, or diagnostic task; only afterwards do they see themselves in a dashboard. In the replication workflow, students first identify what the article makes explicit, what it implies, what it leaves to the implementor, and what remains genuinely ambiguous; only then do they generate code.

This design also connects to the paper's discussion of curiosity and the description-experience gap. The dashboard is not a static report. It is a controlled encounter with incompleteness: the participant sees a gap between what they assumed and what the data show; the reader of an article sees a gap between what the paper claims and what can actually be reconstructed. In both cases, the dashboard is designed to produce a question before supplying an answer.

A.2 Two groups of prompts

The prompt library is divided into two functional groups. The first group contains prompts used for writing and AI-assisted authorship. The second group contains operational prompts for generating dashboards, replication laboratories, and model tools.

A.2.1 Prompts for writing and AI collaboration

The first group of prompts documents how AI was used during writing, editing, and methodological clarification. These prompts do not generate dashboards. They define the author-AI relation.

Their purpose is transparency. They show what kind of assistance the author asked from AI, how the authorial voice was protected, how the argument was disciplined, and how AI was prevented from becoming the source of the research question.

The main files are:

00_META_KNOW_THYSELF.txt
00_META_SLOWACKI_MCCLOSKEY.txt

The Know Thyself meta-prompt defines the intellectual logic of the method: models as stories about human beings, experience before theory, dashboard as mirror, heterogeneity as evidence rather than noise, and AI as infrastructure rather than author. The Słowacki-McCloskey protocol defines the writing relation: the author supplies the ideas, architecture, claims, and responsibility; AI helps express them clearly, economically, and without academic filler.

These prompts are included because AI-assisted writing should not remain hidden. In this project, AI is not treated as an invisible ghostwriter. It is treated as an explicit infrastructure of thinking, editing, and implementation. The intellectual responsibility remains with the author.

A.2.2 Operational prompts for reports, dashboards, and laboratories

The second group of prompts is reusable by readers. These prompts are construction tools. They help transform data, articles, simulations, or model files into executable artefacts: R/Shiny dashboards, replication laboratories, and model-validation tools.

The main files are:

10_STYLE_SHINY_FLEXDASHBOARD.txt
11_STYLE_R_CODE_SAFE_DEPLOYMENT.txt
20_KNOW_THYSELF_AD_HOC_RESEARCH_DASHBOARD.txt
30_REPLICATION_LAB_ARTICLE_v2.txt
40_ABM_SPEC_FROM_NETLOGO.txt
41_ABM_SPEC_FROM_R.txt
42_NETLOGO7_GENERATE_MODEL.txt
43_NETLOGO_VALIDATE_WITH_SHINY_LOGOLINK.txt

These files are not meant to be merged into one master prompt. They form a small prompt library. The style prompts define technical conventions; the task-specific prompts define the kind of artefact to be built.

Table A1. Functional structure of the repository

Prompt family	Function in the method
Writing and AI transparency	Documents how AI supports writing, critique, editing, and conceptual discipline without replacing the authorial question.
Technical style guides	Defines RMarkdown, flexdashboard, Shiny, package loading, HTML, mathematical notation, plotting, and deployment conventions.
Ad hoc research dashboards	Builds dashboards from surveys, classroom experiments, simulations, diagnostic tasks, and short research exercises.
Replication laboratories	Turns scientific articles into interactive laboratories for reconstruction, robustness analysis, and critical reading.
ABM and NetLogo tools	Extracts, generates, documents, and validates agent-based models in NetLogo or R.

A.3 Know Thyself Ad Hoc Research Dashboards

Use the ad hoc research dashboard prompt when the input is a participant-centred instrument: a classroom experiment, survey, simulation, diagnostic task, valuation exercise, short exploratory exercise, or any other situation in which participants first act and then see themselves in the resulting data. Throughout the repository, the term ad hoc research dashboard is deliberately broader than experiment dashboard. The common feature is not experimental control, but the production of a mirror: participants first generate traces of judgement, behaviour, or reconstruction, and only afterwards confront these traces in an interpretable dashboard.

The basic workflow is:

Run the survey, experiment, simulation, or classroom activity.
Export the data or prepare a small sample of results.
Paste the ad hoc research dashboard prompt into the AI assistant.
Paste the sample of results into the same conversation.
Work step by step: identify the data structure, define the dashboard story, build one prototype tab, then extend the dashboard.
Run the resulting .Rmd file locally in RStudio or deploy it as a Shiny/flexdashboard application.

The output is not merely a report. It is a mirror. A participant should be able to ask: Where am I in these data? Was my decision typical? How did my answer differ from the group? What does this reveal about the model's assumptions?

Typical components of an ad hoc research dashboard include:

an opening panel explaining what the participant experienced;
a nickname selector allowing the participant to locate themselves anonymously;
plots showing the participant against the group distribution;
heterogeneity views: distributions, outliers, trajectories, or types rather than only averages;
simple calculators or simulations linking participant behaviour to the theoretical mechanism;
an interpretive sequence that moves from experience to mirror, then to history, theory, and reflection;
a final section with open questions rather than a closed summary.

The dashboard therefore enacts the central claim of the method: theory arrives after the participant has encountered a question in their own data.

A.4 Replication Laboratory Dashboards

Use the replication laboratory prompt when the input is a scientific article. The output is an interactive dashboard that allows students to reconstruct, test, and interrogate the article. The basic workflow is:

Paste the replication laboratory prompt into the AI assistant.
Attach or paste the target scientific article.
Run Stage 1: extract what is explicit, inferred, assumed by the implementor, and genuinely ambiguous.
Review the ambiguity list before allowing any code to be written.
Run Stage 2: design the dashboard architecture.
Run Stage 3: generate the complete .Rmd file.
Run Stage 4: produce an extension guide explaining how new tests or experiments can be added.

The first stage is the methodological core. The ambiguity list identifies where the article is silent but an implementation requires a decision. This is why the protocol forbids code before Stage 1 is complete. Without this constraint, AI may generate technically plausible code that silently resolves the article's underdeterminations, producing a dashboard that looks rigorous but conceals the epistemological problem.

Every replication dashboard must allow the student to ask three questions:

What is certain? What does the article actually establish, not merely suggest?
What is assumed? Where must the implementor choose because the article is silent?
What is fragile? Under which changes in assumptions do the results disappear, reverse, or lose their meaning?

Typical components of a replication laboratory include:

Article Overview: mechanism, research question, data status, and a table distinguishing explicit claims from implementor assumptions.
Baseline Replication: reconstruction of key results and a number-by-number comparison with the article.

Internal Validity: seed sensitivity, initialisation sensitivity, convergence diagnostics, and alternative interpretations of ambiguous points.

External Validity / Sensitivity: parameter sweeps and boundary tests showing where results are robust and where they are fragile.

Specialised Module: simulation model, empirical article with data/code, or empirical article without data/code.

Educational Guide: a map from theory to code, in which important code blocks are explained as modelling choices.

Results and Notes: replication summary, unresolved ambiguities, uncertainty map, and methodological conclusions.

Numerical values in replication laboratories are labelled by evidential status:

[ART] - value taken directly from the article

[IMP] - implementor assumption where the article is silent

[SYN] - synthetic or reconstructed value

This labelling prevents reconstructed or assumed values from acquiring the false authority of reported data. It is also the mechanism by which the dashboard teaches scientific caution.

A.5 Three replication modes

The replication laboratory prompt distinguishes three article types. The same four-stage protocol applies in all three cases, but the specialised module differs.

Module A is used for simulation articles: agent-based models, cellular automata, ODEs, system dynamics, and similar computational models without empirical data. The dashboard focuses on implementation fidelity, seed sensitivity, parameter sweeps, initial conditions, model dynamics, and phase boundaries.

Module B is used for empirical articles with published data and/or code. The dashboard focuses on specification robustness, outlier sensitivity, alternative control sets, subgroup analysis, and coefficient stability across defensible modelling choices.

Module C is used for empirical articles without data or code. The dashboard treats published statistics as constraints on possible datasets. Approximate Bayesian Computation can be used to reconstruct the region of possible datasets consistent with the published tables. Power

analysis and dependency-structure checks help evaluate how informative the reported results actually are.

This division connects the repository to the replication literature and the open-science debate. It is not limited to detecting errors. Its purpose is to teach what can and cannot be known from an article, given the information it provides.

A.6 Code as an educational layer

The generated code is part of the educational artefact. It is not merely an implementation hidden behind the dashboard. A careful reader should be able to inspect the code and understand what is being built.

Every important block of code should explain:

what object is being created;
what assumption it encodes;
which part of the theory, experiment, survey, simulation, or article it corresponds to;
what would change if this block were modified;
how the reader can verify that the code creates what it claims to create.

Comments should not merely describe syntax. They should reveal the mechanism.

A weak comment says:

```
# calculate mean
```

An educational comment says:

```
# We calculate the group mean because the dashboard uses it as
# the benchmark against which each participant can locate their own answer.
# This is the moment where individual experience becomes a social mirror.
```

This principle is especially important for AI-generated code. The code should make the assistant's construction visible: where the data enter, where assumptions enter, where the model is simulated, where results are aggregated, and where interpretation begins.

A.7 Technical conventions

The dashboards use RMarkdown, flexdashboard, and Shiny. The technical prompts define stable conventions for package loading, file paths, plotting, reactivity, HTML,

and mathematical notation. These conventions record implementation knowledge accumulated through repeated debugging.

General conventions include:

full, self-contained .Rmd files;
server-safe package loading through a requiredPackages vector;
base R plotting unless another system is explicitly required;
eventReactive for costly computations;
no automatic re-rendering of plots after every slider change;
project-folder file paths rather than hard-coded external paths;
HTML used primarily for dynamic output;
mathematical notation placed in static RMarkdown whenever possible;
no LaTeX backslashes inside HTML() blocks.

These constraints are not aesthetic. They are part of the method's reproducibility. A dashboard that fails technically cannot function as a mirror, a replication laboratory, or an epistemic provocation.

A.8 ABM and NetLogo tools

The repository also contains prompts for working with agent-based models. These prompts can reverse-engineer an existing NetLogo or R-based model into a formal specification, generate a NetLogo 7 .nlogox model from a specification, or build a validation dashboard for an existing model.

These tools extend the same logic to modelling practice. They make assumptions visible, model dynamics inspectable, and validation experiments reproducible. They are especially useful when a model exists as code but lacks a formal methodological description suitable for teaching or publication.

A.9 Suggested use by readers

Readers who want to build their own dashboard should begin by selecting the appropriate prompt family.

For a survey, classroom experiment, simulation, or diagnostic task, use the ad hoc research dashboard prompt. Paste the prompt and a sample of the results into the chat. First clarify the data structure, then build a single prototype tab, then extend the dashboard.

For a scientific article, use the replication laboratory prompt. Paste the prompt and the target article into the chat. Do not proceed to code until the ambiguity list has been reviewed. The final output should be a complete R/Shiny flexdashboard replication laboratory.

For an existing agent-based model, use the NetLogo or R specification prompt. If the goal is validation, use the NetLogo validation dashboard prompt. If the goal is model construction, use the NetLogo 7 generation prompt.

For all executable dashboards, combine the task-specific prompt with the technical style guide. This ensures consistency in code structure, reactivity, plotting, package loading, file paths, and educational comments.

A.10 Reproducibility and evolution

The repository is versioned because the prompts will evolve. The method develops through classroom use, replication attempts, debugging, and new modelling tasks.

For scholarly purposes, cite the archived release corresponding to this paper. For practical use, consult the current GitHub version. The archived release preserves the exact prompt system used in the paper. The living repository preserves the method's development. Suggested citation:

Kopczewski, T. (2026). Know Thyself Prompt Library: Replication Labs, Ad Hoc Research Dashboards, and ABM Tools. Working repository. https://github.com/tomvar/know-thyself-prompt-library

After the first archived release, replace the temporary citation with the DOI-based citation.

A.11 Boundary

The repository makes a narrower claim. A structured prompt library can encode methodological discipline, technical implementation knowledge, and the sequence of questions required to turn a classroom result or scientific article into an inspectable educational object.

AI lowers the infrastructural cost. It does not supply the question. The human contribution remains the curiosity, errors, hesitations, disagreements, and interpretive judgments that make the dashboard worth building.

University of Warsaw
Faculty of Economic Sciences
44/50 Długa St.
00-241 Warsaw
www.wne.uw.edu.pl
ISSN 2957-0506