

## KRIPKE SEMANTICS, COMMON KNOWLEDGE, BOUNDED RATIONALITY AND AUMANN'S AGREEMENT THEOREM

Dushko Josheski<sup>1\*</sup>, Natasha Miteva<sup>2</sup>, Dushica Popova<sup>3</sup>

<sup>1</sup>Associate professor Faculty of Tourism and Business Logistics, Goce Delcev University, Stip, North Macedonia; e-mail: [dusko.josevski@ugd.edu.mk](mailto:dusko.josevski@ugd.edu.mk) ; ORCID: 0000-0002-7771-7910

<sup>2</sup>Associate professor, Faculty of Tourism and Business Logistics, Goce Delcev University, Stip, North Macedonia, [natasa.miteva@ugd.edu.mk](mailto:natasa.miteva@ugd.edu.mk) ; ORCID: 0009-0004-8082-6458

<sup>3</sup>Assistant professor, Faculty of Tourism and Business Logistics, Goce Delcev University, Stip, North Macedonia, R.North Macedonia, e-mail: [dusica.saneva@ugd.edu.mk](mailto:dusica.saneva@ugd.edu.mk) ; ORCID: 0009-0004-4750-8992

\*Corresponding author: [dusko.josevski@ugd.edu.mk](mailto:dusko.josevski@ugd.edu.mk)

### Abstract

The first model with k-level thinking with Gaussian noise shows that small deviations for common knowledge led to “almost” common knowledge equilibria. The second model demonstrated the semantic economy idea: as agents exchange and adapt beliefs, they create shared informational value by reaching a consensus that reflects network-wide insights rather than mere individual optimization. Higher reasoning depth does not change Nash equilibria but shifts up Kantian beliefs. The shift up in Kantian beliefs suggests a greater alignment toward strategies that maximize collective welfare, rather than purely individualistic or competitive outcomes. This doesn't alter the Nash equilibrium, where players still act independently, but it emphasizes a higher baseline of cooperative or altruistic expectations among players due to more profound belief hierarchies in the reasoning process. In the third model: networked economic context where agents interact in an economic network where competitive advantage depends on the informational value generated across the network, results differ from the second example: Nash beliefs adjust based on others' best responses (shift up), while Kantian beliefs account for mutual benefit, dampening large shifts. Nash and Kantian equilibria differ when only three agents exist versus network economy.

**Keywords:** Kripke semantics, Common knowledge, Kantian equilibrium, Nash equilibrium, Agreement theorem, Bounded rationality

**JEL codes:** C72, C79, D83, D84

### 1.Introduction

In game theory it is more than well known that phenomena such as market speculations and “agree to disagree” can not be observed in equilibrium in a model of Bayesian rational agents, see [Geanakoplos, J. \(1989/2021\)](#). In classical game theory, or the neoclassical school of economic thought, rationality is assumed on the assumption that agents can unilaterally change their strategies, so [Stalnaker\(1994\)](#) applied [Kripke \(1963\)](#) work to game theory and showed that rationalizability is characterized with common belief in rationality, and that Nash equilibria are characterized with rationality and knowledge of the opponents belief, see [Fourny, G. \(2018\)](#) .But Nash concept seemed weak in some ways ,too strong in others, to “yield plausible recommendations in all cases”, see [Stalnaker\(1994\)](#). [Myerson \(1991\)](#), expresses doubt that any solution concept can satisfy all criteria of adequacy<sup>1</sup>, suggesting that perhaps

---

<sup>1</sup> For instance, does Nash solution concept satisfy general existence theorem? The answer is that the two concepts differ: Nash Equilibrium is a concept from non-cooperative game theory, where players (agents) make individual decisions with the goal of maximizing their own payoffs, given the actions of others. The General Existence Theorem pertains to competitive markets and the existence of a general equilibrium in an economy

the best we can do is to find various notions that offer a useful way "to formalize part of our intuitive criteria about how rational intelligent players might behave in a game." So we introduced the concept of Kantian equilibrium in our analysis, following [Osborne.M.J., Rubinstein,A.\(2023\)](#), and [Roemer \(2010\)](#), [Roemer \(2019\)](#). Similar to the General Existence Theorem, Kantian equilibria tend to result in Pareto-efficient outcomes since players' strategies maximize collective welfare. This parallels the goal of the General Existence Theorem by [Arrow, Debreu \(1954\)](#), which is to achieve efficient allocation through market equilibrium. But for the existence of equilibria: For Kantian equilibria, existence theorems are often proven using fixed-point theorems like Brouwer's or Kakutani's, similar to Nash equilibrium proofs, rather than the Walrasian conditions required by the General Existence Theorem<sup>2</sup>. Knowledge and interactive knowledge are central elements in economic theory. So again, this question can rational agents "agree to disagree"? see [Geanakoplos \(1992\)](#). Bayesian Nash equilibrium<sup>3</sup> implies that agents cannot agree to disagree; it implies that they cannot bet when the bet is common knowledge; and most surprising of all, it does not include speculation. So here comes no trade theorem The agreement theorem underlies an important set of results that place limits on the trades that can occur in differential information models under the common prior assumption(see [Kreps, \(1977\)](#); [Milgrom and Stokey, \(1982\)](#)). These no-trade theorems(The No-Trade Theorem is an economic theory that explains why rational agents with common knowledge and identical prior beliefs should not engage in trade purely based on the exchange of information) state, in various ways, that rational risk-averse traders cannot take opposite sides on a purely speculative bet. Common knowledge is one of the key components of no-trade theorem. Common knowledge means that all agents not only know the available information but also know that others know this information—and they know that everyone else knows that everyone knows, and so on (infinite recursion). In the context of the No-Trade Theorem, if there's common knowledge about asset values, market conditions, or other relevant factors, no agent can gain a unique informational advantage. The theorem assumes that agents start with identical prior beliefs about the state of the world and update these beliefs in a Bayesian manner upon receiving new information. If agents start with the same beliefs and interpret information in the same way, any information revealed by one agent's desire to trade will immediately lead others to adjust their beliefs, again removing the incentive for trade. [Rubinstein \(2021\)](#) proposes 4 models of bounded rationality: Limited ability to solve a set of propositions, Reducing the complexity of strategies, Belief formation on the basis of a small sample; Diversified views of the world. In this paper we include Luce choice axiom [Luce, R. D. \(1959/2005\)](#) with bounded rationality which is rationality in a sense of [Selten \(1998\)](#).,[Simon \(1957\)](#). These were all motivations for writing this paper. Kripke semantics was used in modeling to represent possible worlds and accessibility, and knowledge representation (For any agent  $i$  in world  $w$ , the agent knows  $p$  if  $p$  holds in all worlds accessible to  $i$  from  $w$ ). Bounded rationality was represented through  $k$ -level thinking—where each agent reasons only up to a finite depth, limiting how much they can consider other agents' potential reasoning. Aumann's Agreement theorem when applied to agents in a Kripke frame, common knowledge is achieved through repeated information exchanges. Cognitive Hierarchy Theory describes agents who operate at different levels of reasoning (or depth) in our models  $K$ -Level thinking captures this by allowing each agent to compute beliefs based on others' beliefs

---

where prices adjust so that supply equals demand across all markets, see [Arrow, Debreu \(1954\)](#); [McKenzie, \(1954\)](#).

<sup>2</sup> Kantian equilibrium does not rely on prices to coordinate agents' actions; instead, it relies on a shared cooperative principle. Kantian equilibrium does not involve clearing of markets as in a Walrasian equilibrium, which is central to the General Existence Theorem. Kantian equilibrium can apply to cooperative settings where players are not necessarily price-takers, which is a core assumption of the General Existence Theorem unlike Nash equilibria.

<sup>3</sup> Common knowledge of rationality and optimization

iteratively up to their reasoning depth. And semantic economy goals are achieved through cooperation, knowledge-sharing, and creating shared understandings.

## 2.Kripke semantics,classical and intuitionistic Kripke model, and Zermelo-Fraenkel set theory

Kripke semantics is a formal system used to model modal logic—logics that involve modalities like necessity ( $\Box$ ) and possibility ( $\Diamond$ ). These modalities express statements about what is necessarily true or possibly true, often in the context of knowledge, belief, or time. Kripke semantics provides a structure to interpret these modalities using possible worlds. The concept of Kripke model is due to [Kripke \(1959\)](#), [Kripke\(1962\)](#),[Kripke \(1963\)](#),[Kripke \(1965\)](#).

*Definition 1* A classical Kripke model (due to [Ilik et al.\(2010\)](#)) is given by a quintuple  $(K, \leq, D, \Vdash_s, \Vdash_\perp)$ ,  $K$  inhabited, such that  $(K, \leq)$  is a poset<sup>4</sup> of possible worlds,  $D$  is the domain function assigning sets to the elements of  $K$  such that:

*Well-formed formula 1*

$$\forall w, w' \in K, (w \leq w' \Rightarrow D(w) \subseteq D(w'))$$

i.e.,  $D$  is monotone. Let the language be extended with constant symbols for each element of  $\mathcal{D} := \cup \{D(w) : w \in K\}$ . And,  $(-): (-) \Vdash_s$  is a binary relation of “strong refutation” between worlds and atomic sentences in the extended language such that:

*Well-formed formula 2*

$$\begin{aligned} w: X(d_1, \dots, d_n) \Vdash_s &\Rightarrow d_i \in D(w), \forall i \in \{1, \dots, n\} \\ w: X(d_1, \dots, d_n) \Vdash_s, w \leq w' &: X(d_1, \dots, d_n) \Vdash_s \end{aligned}$$

The relation  $\Vdash$  is called the *satisfaction relation, evaluation, or forcing relation*<sup>5</sup> Given a model  $M$  (usually a transitive model of ZFC-Zermelo–Fraenkel set theory, see [Zermelo \(1930\)](#)), any poset  $(P, <)$  in it is a notion of forcing and its elements forcing conditions. A  $G$  in  $M$  is said to be generic if it is a filter and any dense set in  $P$  that belongs to  $M$  has a nonempty intersection with  $G$ . There's a theorem that states that for a transitive model  $M$  of ZFC and a generic set  $G \subset P$  there's a transitive model  $M[G]$  of ZFC that extends  $M$  and, associated with that, we define a forcing relation  $\Vdash$  where some element  $p \in G$  forces a formula  $\varphi$  iff  $M[G] \models \varphi$ , i.e.,  $(\exists p \in G) p \Vdash \varphi$  if  $\varphi$  is valid in  $M[G]$ , this will happen for every generic  $G$  if  $\varphi$  is said to be in the forcing language.

*Definition 2* For  $\mathbb{P} \in V$  a poset and  $p \in \mathbb{P}$ , we say  $p$  forces  $\varphi$  and write  $p \Vdash \varphi$  iff for every generic over  $V$  filter  $X$  containing,  $p$ ,  $V[X] \models \varphi$ .

*Definition 3* The relation  $(-): (-) \Vdash_s$  of strong refutation is extended to the relation between worlds  $w$  and composite sentences  $A$  in the extended language with constants in  $D(w)$ , inductively, together with the two new relations. A sentence  $A$  is forced in the world  $w$  (notation  $w : A \Vdash$ ) if any world  $w' \geq w$ , which strongly refutes  $A$ , is exploding. A sentence  $A$  is forced in the world  $w$  (notation  $w : A \Vdash$ ) if any world  $w' \geq w$ , which forces  $A$ , is exploding

<sup>4</sup> A partially ordered set (normally, poset) is a set,  $L$ , together with a relation,  $\leq$ , that obeys,  $\forall a, b, c \in L$ : (reflexivity)  $a \leq a$ ; (anti-symmetry) if  $a \leq b$  and  $b \leq a$  then  $a = b$ ; and (transitivity) if  $a \leq b$  and  $b \leq c$  then  $a \leq c$ . The relation  $\leq$  is called a partial order on  $L$ . See, [Dickson \(2007\)](#).

<sup>5</sup> In mathematics or set theory forcing is a technique for proving consistency and independence results.

Well-formed formula 3

$$\begin{aligned}
 w: A \wedge B \Vdash_s & \text{ if } w: A \Vdash \bigvee w: B \Vdash \\
 w: A \vee B \Vdash_s & \text{ if } w: A \Vdash \bigwedge w: B \Vdash \\
 w: A \rightarrow B \Vdash_s & \text{ if } w: \Vdash A \bigwedge w: B \Vdash \\
 w: \forall x A(x) & \text{ if } w: A(d) \Vdash \text{ for some } d \in D(w) \\
 w: \exists x. A(x) \Vdash_s, & \forall w' \geq w \wedge d \in D(w'), w': A(d) \Vdash; \\
 \perp & \text{ is always strongly refuted} \\
 \top & \text{ is never strongly refuted}
 \end{aligned}$$

The Zermelo-Fraenkel axioms are the basis for Zermelo-Fraenkel set theory.

1. *Axiom of Extensionality*: If  $X$  and  $Y$  have the same elements, then  $X = Y$ .

Well-formed formula 4

$$\forall u (u \in X \equiv u \in Y) \Rightarrow X = Y$$

2. *Axiom of the Unordered Pair* (axiom of pairing): For any  $a$  and  $b$  there exists a set  $\{a, b\}$  that contains exactly  $a$  and  $b$ .

Well-formed formula 5

$$\forall a \forall b \exists c \forall x \left( x \in c \equiv (x = a \vee x = b) \right)$$

3. *Axiom of subsets* (*Axiom of Separation* or *Axiom of Comprehension*): If  $\varphi$  is a property (with parameter  $p$ ), then for any  $X$  and  $p$  there exists a set  $Y = \{u \in X : \varphi(u, p)\}$  that contains all those that have the property  $\varphi$ .

4. *Axiom of the sum of set* (*Axiom of Union*): For any  $X$  there exists a set  $Y = \cup X$ , the union of all elements of  $X$ .

Well-formed formula 6

$$\forall X \exists Y \forall u (u \in Y \equiv \exists z (z \in X \wedge u \in z))$$

5. *Axiom of the power set*: For any  $X$  there exists a set  $Y = P(X)$ , the set of all subsets of  $X$ .

Well-formed formula 7

$$\forall X \exists Y \forall u (u \in Y \equiv u \subseteq X)$$

6. *Axiom of Infinity*: There exists an infinite set.

Well-formed formula 8

$$\exists S \left[ \emptyset \in S \wedge (\forall x \in S) \left[ x \cup \{x\} \in S \right] \right]$$

7. *Axiom of Replacement*: If  $F$  is a function, then for any  $X$  there exists a set  $Y = F[X] = \{F(x) : x \text{ in } X\}$ .

Well-formed formula 9

$$\forall x \forall y \forall z \left[ \varphi(x, y, p) \wedge \varphi(x, z, p) \Rightarrow y = z \right] \Rightarrow \forall X \exists Y \forall y \left[ y \in Y \equiv (\exists x \in X) \varphi(x, y, p) \right]$$

8. *Axiom of Foundation*: Every nonempty set has an  $\in$  in -minimal element.

*Well-formed formula 10*

$$\forall S \left[ S \neq \emptyset \Rightarrow (\exists x \in S) S \cap x = \emptyset \right]$$

9. *Axiom of Choice*: Every family of nonempty sets has a choice function.

*Well-formed formula 11*

$$\forall x \in a \exists A(x, y) \Rightarrow \exists y \forall x \in a A(x, y(x))$$

A *Kripke model*  $\mathcal{M}$  for a propositional logical system  $\Lambda$ , classical, intuitionistic, or modal is a pair  $(\mathcal{F}, V)$ , where  $\mathcal{F} = (W, R)$  is a Kripke frame, and  $V$  is a function that takes each atomic formula of  $\Lambda$  to a subset of  $W$ . If  $w \in V(p)$ , we say that  $p$  is *true* at world  $w$ . We say that  $\mathcal{M}$  is a  $\Lambda$ -model *based on* the frame  $\mathcal{F}$  if  $\mathcal{M} = (\mathcal{F}, V)$  is a model for the logic  $\Lambda$ . Since the well-formed formulas (wff's) of  $\Lambda$  are uniquely readable,  $V$  may be inductively extended so it is defined on all wff's. The following are some examples:

*Well-formed formula 12*

in classical propositional logic:  $PL_c V(A \rightarrow B) := V(A)^c \cup V(B); S^c := W - S$

in the model propositional logic:

*Well-formed formula 13*

$$K, V(\Box A) := V(A)^\Box, \text{ where } S^\Box := \{u \mid \uparrow u \subseteq S\}, \wedge \uparrow u := \{w \mid uRw\}$$

In Propositional Intuitionistic Logic:

*Well-formed formula 14*

$$P_i V(A \rightarrow B) := (V(A) - V(B))^\#, S^\# := (\downarrow S)^c, \wedge \downarrow S := \{u \mid uRw, w \in S\}$$

About Kripke semantics, A Kripke model is a triple  $\mathcal{M} = \langle W, R, v \rangle$ , where  $W$  a non-empty set of possible worlds,  $R$  is a preorder (i.e., a reflexive and transitive relation) on  $W$ , and  $v: Var \times W \rightarrow \{0, 1\}$  is the *variable valuation function*. The function  $v$  is required to be monotonic w.r.t.  $R$ : if  $xRy$ , then  $v(p, x) \leq v(p, y) \forall p \in Var$ . In other words, if  $v(p, x) = 1$  and  $xRy$ , then  $v(p, y) = 1$ . By  $R(x)$  we denote the set  $\{y \mid xRy\}$ . In different worlds, different formulae are considered true. If formula  $A$  is true in world  $x$  of  $\mathcal{M}$ , we write  $\mathcal{M}, x \Vdash A$ ; is called the forcing relation and defined as follows:

*Well-formed formula 15*

$$\begin{aligned} &\mathcal{M}, x \not\Vdash \perp, \text{ falsity is never true} \\ &\mathcal{M}: x \Vdash p, \text{ iff } v(p, x) = 1 \\ &\mathcal{M}: x \Vdash A \wedge B \text{ iff } \mathcal{M}, x \Vdash A, \mathcal{M}, x \Vdash B \text{ (conjunction)} \\ &\mathcal{M}: x \Vdash A \vee B \text{ iff } \mathcal{M}, x \Vdash A, \mathcal{M}, x \Vdash B \\ &\mathcal{M}: x \Vdash A \rightarrow B \text{ iff } \forall y \in R(x), \mathcal{M}, y \not\Vdash A, \wedge \mathcal{M}, y \Vdash B \end{aligned}$$

This definition is designed to preserve monotonicity of forcing: if  $\mathcal{M}, x \Vdash A$  and  $xRy$ , then  $\mathcal{M}, y \Vdash A$ . If the Kripke model has only one world ( $|W| = 1$ ), then it is a model for classical propositional logic. Intuitionistic propositional logic is sound w.r.t. Kripke semantics:

*Theorem 1* if  $\vdash_{Int} A$ , then for every Kripke model  $\mathcal{M} = \langle W, R, v \rangle$  and for every possible world  $x \in W$  of this model  $\mathcal{M}, x \Vdash A$ .

*Proof:* in order to prove soundness, one needs to prove that if  $A$  is an axiom of Int, then  $\mathcal{M}, x \Vdash A$  and second if  $\mathcal{M}, x \Vdash A \rightarrow B \Rightarrow \mathcal{M}, x \Vdash B$ . The second part is easy: If  $x \Vdash A \rightarrow B$ , then for every world  $y \in R(x)$  we have either  $y \not\Vdash A$  or  $y \Vdash B$ . Since  $y = x$  by reflexivity of  $R$ , then given  $x \Vdash A$ , we obtain  $x \Vdash B$ . Here, we need to prove  $x \Vdash A \rightarrow (B \rightarrow C) \rightarrow ((A \rightarrow B) \rightarrow (A \rightarrow C))$ . In order to establish that a formula of the form  $E \rightarrow F$  is true in  $x$ , one needs to check that  $\forall y \in R(x)$  if  $y \Vdash E$ , then  $y \Vdash F$ . Again, let's consider arbitrary  $z \in R(y)$ , such that  $z \Vdash A \rightarrow B$ . On this turn we need to show that  $z \Vdash A \rightarrow C$ . Let  $w$  be a world from  $R(z)$ , such that  $w \Vdash A$  and finally we need  $w \Vdash C$ . So, now:

*Well-formed formula 16*

$$\begin{array}{c} w \Vdash A \\ \uparrow \\ z \Vdash A \rightarrow B \\ \uparrow \\ y \Vdash A \rightarrow (B \rightarrow C) \\ \uparrow \\ x \end{array}$$

By monotonicity, since  $yRw, zRw$ , the formulae  $A \rightarrow (B \rightarrow C)$  and  $A \rightarrow B$  are also true in  $w$ . Since modus ponens<sup>6</sup> is applicable for  $\Vdash$ , we have  $w \Vdash B \rightarrow C, w \Vdash B, w \Vdash C$  which is our goal ■.

Now, about Kripke *completeness theorem*.

*Theorem 2* If a formula is true in every possible world of any Kripke model, then it is derivable in Int.

We proceed, let  $A$  be the formula such that  $\not\Vdash_{Int} A$ . Now, a countermodel for  $A$ , that is model  $\mathcal{M}$  that contains a world  $x$ , such that  $\mathcal{M}, x \not\Vdash A$ . This will be the canonical model for Int, denoted by  $\mathcal{M}_0$ .

*Definition 4* A set  $\Gamma$  of formulae is called disjunctive theory<sup>7</sup>:

1.  $\Gamma$  is deductively closed, i.e.  $\Gamma \vdash_{Int} B$ , so  $B \in \Gamma$
2.  $\Gamma$  is consistent i.e.  $\Gamma \not\Vdash_{Int} \perp$
3.  $\Gamma$  is disjunctive i.e.  $\Gamma \not\Vdash_{Int} A \vee B, \Gamma \vdash_{Int} A \wedge \Gamma \vdash_{Int} B$

*Definition 5* The *canonical model* for Int is the model  $\mathcal{M}_0 = \langle W_0, R_0, v_0 \rangle$  where:  $W_0$  is the set of all disjunctive theories,  $R_0$  is the subset relation  $\Gamma R_0 \Gamma_2 \Leftrightarrow \Gamma_1 \subseteq \Gamma_2, v_0$  is defined as follows:  $v_0(p, \Gamma) = 1 \Leftrightarrow p \in \Gamma$ .

*Lemma 1* Let  $\mathcal{M}_0, \Gamma \Vdash B \Leftrightarrow B \in \Gamma$

This is called *Main Semantic Lemma*. Or the Main Semantic Lemma states:

1. If a formula  $\phi$  is provable in the modal logic i.e.  $\vdash \phi$ , then  $\phi$  is true in all models i.e.  $M \vDash \phi$ .

<sup>6</sup> It can be summarized as "P implies Q. P is true. Therefore, Q must also be true." Or  $\frac{P \rightarrow Q, P}{Q}$  see [Stone \(1996\)](#).

<sup>7</sup> Disjunctive theory typically refers to a theory or logical framework that uses disjunctions (logical OR statements) as a central component. The disjunction is a fundamental logical connective in both classical and non-classical logics, and a disjunctive theory would emphasize the role of such disjunctions in reasoning or inference.

2. If  $\phi$  is true in all models, then  $\phi$  is provable i.e.  $M \models \phi$  implies  $\vdash \phi$ .

For this part of the paper see more in <https://homepage.mi-ras.ru/~sk/lehre/penn2017/> Logic II (LGIC 320 / MATH 571) (University of Pennsylvania, Spring 2017).

*Lemma 2* This holds in classical Kripke semantics

1.  $w: \Vdash A \Leftrightarrow \neg A \Vdash_S$
2.  $w: A \Vdash \Leftrightarrow w: \Vdash \neg A$
3.  $w: \neg A \Vdash \Leftrightarrow w: \Vdash A$
4.  $w: \neg A \Vdash \Leftrightarrow w: \neg A \Vdash_S$
5.  $w: \Vdash A \Leftrightarrow w: \Vdash \neg \neg A$
6.  $w: A \Vdash \Leftrightarrow w: \Vdash \neg \neg A \Vdash$
7.  $w: \neg A \Vdash_S \Leftrightarrow w: \Vdash \neg \neg A \Vdash \Leftrightarrow w: \Vdash A$

*Proof:* under number 1 obvious because  $w: \perp \Vdash$ ; under second it is obvious because:  $w: \Vdash A \rightarrow B \Leftrightarrow \forall w' \geq w, w': \Vdash A \Rightarrow w': \Vdash B$ ,  $w: \Vdash A \wedge B \Leftrightarrow w: \Vdash A \wedge w: \Vdash B$ ,  $w: \Vdash A \vee B \Leftrightarrow w: \Vdash A \vee w: \Vdash B$ ,  $w: \Vdash A \vee B \Leftrightarrow w: \Vdash A \vee w: \Vdash B$ ,  $w: \Vdash \exists x A(x) \Leftrightarrow \forall d \in D(w), w: \Vdash A(d)$ ■, see [Ilik et al.\(2010\)](#).

So, now in turn basic elements of Kripke frame can be simply written as:

*Well-formed formula 17*

$$\mathcal{F} = (W, R)$$

$W$  is non-empty set of worlds,  $R \subseteq W \times W$  is a **binary accessibility relation** on  $W$ , which determines which worlds are accessible from other worlds.  $wRw'$  means that the world  $w'$  is accessible from world  $w$ . A Kripke model  $\mathcal{M}$  extends the frame:

*Well-formed formula 18*

$$\mathcal{M} = (W, R, v)$$

$W, R$  are the same as in Kripke frame.  $V: Prop \rightarrow 2^W$  is a valuation function, where  $Prop$  is a set of propositional variables<sup>8</sup> and  $2^W$  is the power set of  $W$ . For each propositional variable  $p \in Prop, V(p) \subseteq W$ .

## 2.1 Truth in Kripke Models

The truth of a formula of a world  $w \in W$  in a Kripke model  $\mathcal{M} = (W, R, v)$  is defined inductively. Let  $\varphi$  represents formula in modal logic. Here, we define  $\mathcal{M}, w \models \varphi$  to mean that  $\varphi$  is true in model  $\mathcal{M}$  at world  $w$ . The truth conditions are as follows:

1. For a propositional variable  $p$ :

*Well-formed formula 19*

$$\mathcal{M}, w \models p, \text{ if and only if } w \in V(p)$$

---

<sup>8</sup> A propositional variable in propositional logic is a countable infinite set of symbols denoted by  $V$ , representing unknown truth values that can be either true or false in logical expressions.

That is  $p$  is true at world  $w$  if  $w$  is in the set of worlds where  $p$  is true.

## 2. Negation:

*Well-formed formula 20*

$$\mathcal{M}, w \models \neg\varphi \text{ if and only if } \mathcal{M}, w \not\models \varphi$$

The  $\neg\varphi$  is true at  $w$  if  $\varphi$  is not true at  $w$ .

## 3. Conjunction

*Well-formed formula 21*

$$\mathcal{M}, w \models \varphi \wedge \psi \text{ if and only if } \mathcal{M}, w \models \varphi, \mathcal{M}, w \models \psi$$

That is  $\varphi \wedge \psi$  is true at  $w$  if both  $\psi$  and  $\varphi$  are true at  $w$ .

## 4. Modal operators

For the necessity operator  $\Box\varphi$  :

*Well-formed formula 22*

$$\mathcal{M}, w \models \Box\varphi \text{ if and only if } \forall w' \in W, (wRw' \Rightarrow \mathcal{M}, w' \models \varphi)$$

That is  $\Box\varphi$  is true at  $w$  if  $\varphi$  is true in all worlds  $w'$  that are accessible from  $w$ .

For the possibility operator  $\Diamond\varphi$  :

*Well-formed formula 23*

$$\mathcal{M}, w \models \Diamond\varphi \text{ if and only if } \exists w' \in W, (wRw' \wedge \mathcal{M}, w' \models \varphi)$$

That is,  $\Diamond\varphi$  is true at  $w$  if  $\varphi$  is true in at least one world,  $w'$  that is accessible from  $w$ .

## 2.2 Properties of the Accessibility Relation $R$

Different modal logics impose different conditions on the accessibility relation  $R$ . Some of the key properties are:

1. **Reflexivity:**  $\forall w \in W, wRw$  This means that each world can access itself. Reflexivity corresponds to **knowledge** logic, where if something is true in a world, the agent knows that it is true.
2. **Symmetry:**  $\forall w, w', w'' \in W, wRw' \Rightarrow w'Rw$ . This means that if  $w'$  is accessible from  $w$ , is also accessible from  $w'$ . Symmetry is relevant in **shared knowledge**.
3. **Transitivity:**  $\forall w, w', w'' \in W, (wRw' \wedge w'Rw'') \Rightarrow wRw''$ . This means that if  $w'$  is accessible from  $w$ , and  $w''$  is accessible from  $w'$ . Transitivity is often epistemic logic.
4. **Euclidean:**  $\forall w, w', w'' \in W, (wRw' \wedge wRw'') \Rightarrow w'Rw''$ . This is another condition in epistemic logic.

## 2.3 Validity and Satisfaction in Kripke Models

A formula  $\varphi$  in a Kripke model  $\mathcal{M} = (W, R, v)$  is valid if :

*Well-formed formula 24*

$$\mathcal{M}, w \models \varphi, \forall w \in W$$



This is,  $\varphi$  is true in all possible worlds of the model. A formula  $\varphi$  is valid in Kripke frame  $\mathcal{F} = (W, r)$  if it is true in every Kripke model based on that frame.

### 3. Common knowledge

First, we will outline [Geanakopolos \(1992\)](#) model of common knowledge. Model outline is as follows: Let there be a set  $N$  of agents, where each agent  $i \in N$  holds information. Let the state of the world be represented by  $\Omega$ , which is a set of possible states  $\omega \in \Omega$ . Each agent  $i$  is associated with an information partition  $\mathcal{P}_i$  which is a partition of  $\Omega$ . This partition represents the agent's knowledge, i.e., what states of the world the agent can distinguish. If two states  $\omega$ , and  $\omega'$  are in the same element of  $\mathcal{P}_i$ , then agent  $i$  cannot distinguish between these two states. Now, knowledge can be represented as set theoretic concept, for each agent  $i$ , the information partition  $\mathcal{P}_i$  induces a **knowledge operator**  $K_i$ , where for any event  $E \subseteq \Omega$ ,  $K_i(E)$  is the set of states in which agent  $i$  knows that  $E$  has occurred. Formally we define this as:

equation 1

$$K_i(E) = \{\omega \in \Omega \mid \forall \omega' \in \mathcal{P}_i(\omega), \omega' \in E\}$$

In words agent  $i$  knows that event  $E$  occurs if, at state  $\omega$ , all the states indistinguishable from  $\omega$  i.e. those in same partition cell are also in  $E$ . Common knowledge among all agents can be derived using set theory.

*Definition 6* We can define **common knowledge** of an event  $E$  as the event where everyone knows  $E$ , everyone knows that everyone knows  $E$ , and so on ad infinitum.

This is captured by the common knowledge operator  $K^*(E)$  which is the intersection of all iterated knowledge operators:

equation 2

$$K^*(E) = \bigcap_{n=1}^{\infty} \left( \bigcap_{i_1, i_2, \dots, i_n \in N} K_{i_1}, K_{i_2}, \dots, K_{i_n}(E) \right)$$

This intersection represents the set of states where  $E$  is common knowledge—i.e., where all agents know  $E$ , all agents know that all agents know  $E$ , and so on. Geanakoplos's work often involves Bayesian updating, where agents revise their beliefs based on new information. In a set-theoretic framework, we can model this as follows. Each agent  $i$  has a prior belief which is represented by probability distribution  $\mu_i \in \Omega$ . When agent  $i$  observes an event  $E$ , they update their belief using Bayes' rule. The updated belief  $\mu_i(E|\omega)$  is defined as :

equation 3

$$\mu_i(E|\omega) = \frac{\mu_i(E \cap \mathcal{P}_i(\omega))}{\mu_i(\mathcal{P}_i(\omega))}$$

#### 3.1 Common knowledge in Kripke frame

Kripke frame with valuation function is:

equation 4

$$M = \{W, (R_i)_{i \in N}, V\}$$

$V: Prop \rightarrow 2^W$  is a valuation function that assigns a set of worlds to each proposition  $p \in Prop, \forall p, V(p) \subseteq W$  set of worlds where  $p \rightarrow true$ . Knowledge operator that we should define is  $K_i$  for each agent  $i$ . Now, given a Kripke model  $M = \{W, (R_i)_{i \in N}, V\}$  and a world  $w \in W$ , agent

$i$  knows  $p$  at world  $w$ , if for  $\forall w' \Rightarrow wR_i w'$  i.e.  $w'$  is possible according to agent  $i$ 's knowledge in world  $w$ ,  $p$  holds in  $w'$  or formally:

*Well-formed formula 25*

$$M, w \models K_i p \Leftrightarrow \forall w' \in W, (wR_i w') \Rightarrow M, w' \models p$$

This means that agent  $i$  knows  $p$  at world  $w$  if in all worlds they consider possible  $p \rightarrow true$ . We can define common knowledge<sup>9</sup> here by using iterated knowledge operator over the agents. Now, let  $K_i$  denote the knowledge operator for agent  $i$  and let  $N$  be the set of all agents. The common knowledge operator can be defined recursively:

*equation 5*

$$C_p = \bigcap_{n=1}^{\infty} K_1, K_2, \dots, K_n p$$

Alternatively, we can define common knowledge by creating a new relation  $R_C$  called common knowledge relation, which is transitive closure of the union of the individual relations  $R_i$ :

*equation 6*

$$R_C = \bigcap_{i \in N} R_i$$

Then the common knowledge operator  $C$  can be defined as:

*Well-formed formula 26*

$$M, w \models C_p \Leftrightarrow \forall w' \in W, (wR_C w') \Rightarrow M, w' \models p$$

This means that common knowledge of  $p$  hold at world  $w$  if,  $\forall w'$  worlds that are reachable through the common knowledge relation  $R_C$ ,  $p$  holds:

### 3.2 Set-theoretic model of common knowledge in Kripke semantics

The set of all possible worlds  $W$  is forming the basic structure. The accessibility relations  $R_i \subseteq W \times W$  represents the knowledge of each agent,  $i$ . For a proposition  $p$ , we define set

*equation 7*

$$\llbracket p \rrbracket = \{w \in W \mid M, w \models p\}$$

The common knowledge of  $p$  is the set of worlds where  $p$  holds in all possible worlds that are reachable through the common knowledge relation  $R_C$ :

*equation 8*

$$\llbracket C_p \rrbracket = \{w \in W \mid \forall w' \in W, (wR_C w') \Rightarrow w' \in \llbracket p \rrbracket\}$$

Here the set of worlds  $p$  is a common knowledge, i.e. everyone knows  $p$ , everyone knows that everyone knows  $p$ , and so on. Common knowledge has several important properties, especially in Kripke semantics:

1. **Monotonicity:** if  $p$  is common knowledge, and  $p \Rightarrow q$ , then  $q$  is common knowledge

<sup>9</sup> A very basic assumption of studies in game theory is that the game is common knowledge, see [Rubinstein \(1989\)](#). Situations without common knowledge are labeled as games with incomplete information see [Harsanyi \(1967\) part I](#), [Harsanyi \(1968\) part II](#), [Harsanyi \(1968\) part III](#).

2. **Fixed Point:** Common knowledge of a proposition is a fixed point in the knowledge structure. Once a proposition becomes common knowledge, it remains so.

### 3.3 Iterated knowledge

Common knowledge can also be viewed as limit of iterated knowledge. First, we will define **iterated knowledge operators**  $K^n$ , for  $n \geq 1$  as follows:

equation 9

$$\begin{aligned} K^1 p &= \bigwedge_{i \in N} K_i p \\ K^2 p &= \bigwedge_{i \in N} K_i K^1 p \\ K^3 p &= \bigwedge_{i \in N} K_i K^2 p \end{aligned}$$

$K^1 p$  everyone knows  $p$ ,  $K^2 p$  everyone knows that everyone knows  $p$ ,  $K^3 p$  everyone knows that everyone knows that everyone knows  $p$ . So common knowledge is the limit of the process:

equation 10

$$C_p = \lim_{n \rightarrow \infty} K^n p$$

The knowledge operator  $K_i$  for agent  $i$  can be represented as set operation:

equation 11

$$K_i \llbracket p \rrbracket = \{w \in W \mid \forall w' \in W, (wR_i w') \Rightarrow w' \in \llbracket p \rrbracket\}$$

The set contains all worlds where agent  $i$  knows that  $p$  is true. Common knowledge can be defined as intersection of iterated knowledge sets. The set of worlds where  $p$  is common knowledge is the fixed point of the iterated knowledge process:

equation 12

$$\llbracket C_p \rrbracket = \bigcap_{n=1}^{\infty} K^n \llbracket p \rrbracket$$

In formal logic, **common knowledge** is often described as the **fixed point** of a knowledge process. The **fixed-point theorem** states that common knowledge is the smallest set that satisfies the property of being known by all agents at all iterations. Mathematically, this can be expressed as:

equation 13

$$\llbracket C_p \rrbracket = \{w \in W \mid w \in K_i(\llbracket C_p \rrbracket) \forall i \in N\}$$

### 3.4 Standard model of knowledge as in [Hintikka \(1962\)](#) per [Rubinstein \(1998\)](#)

An information structure is  $(\Omega, P)$ , where  $\Omega$  is a set of states. It is a “full description of world” or at least relevant facts about the world. The second component is a function  $P$  that assigns each state  $\omega$  a non-empty subset of states,  $P(\omega)$ . The assumption that  $P(\omega) \neq \emptyset$  means that the decision maker can not be so “wrong” as to exclude all possible states as being feasible, see [Rubinstein\(1998\)](#). Three properties of information structures usually associated with the term rationality are:

*Property 1*  $\omega \in P(\omega)$

This property expresses the condition that the decision maker never excludes the true state from the set of feasible states. This property ensures that each state  $\omega \in \Omega$  belongs to its own information set  $P(\omega)$ . Hence, for each  $\omega$ , there exists a set  $\exists P(\omega)$  that contains  $\omega$ . This is the **truthfulness** condition: the agent knows the true state is part of their information set.

*Property 2* if  $\omega' \in P(\omega) \Rightarrow P(\omega') \subseteq P(\omega)$

So, it is not possible for a decision maker who satisfies previous property to hold the view that  $\omega' \in P(\omega)$ , despite there being a state  $z$ , so that  $z \in P(\omega')$  and  $z \notin P(\omega)$ . Then, at  $\omega$  a rational decision maker could consider that: the state  $z$  is excluded and the state  $\omega'$  one will not exclude  $z$ . Thus it must be that state is not  $\omega'$ . This contradicts assumption that  $\omega' \in P(\omega)$ . This property implies **positive introspection**, meaning that if the agent considers  $\omega'$  possible when the true state is  $\omega$ , then their information at state  $\omega'$  cannot reveal more information than what they already knew at  $\omega$ . Formally, the information set  $P(\omega')$  at  $\omega$  must be a subset of  $P(\omega)$ .

*Property 3* if  $\omega' \in P(\omega) \Rightarrow P(\omega') \supseteq P(\omega)$

If an information structure satisfies propositions 1,3 also satisfies 2. And, if  $\omega' \in P(\omega)$  then by proposition 3  $P(\omega') \supseteq P(\omega)$ , by proposition 1  $\omega \in P(\omega)$ , and thus  $\omega' \in P(\omega)$ , which by proposition 3 implies that  $P(\omega') \supseteq P(\omega)$ . This property implies **negative introspection**, meaning that if the agent considers  $\omega$  possible when the true state is  $\omega$ , then the information set  $P(\omega')$  must reveal at least as much information as  $P(\omega)$ . Formally, the information set  $P(\omega')$  must contain  $P(\omega)$ .

*Proposition 1* An information structure  $\Omega, P$  is partitional if and only if it satisfies Properties 1,2,3.

*Proof:*

**1. Property 1 (Truthfulness):**  $\omega \in P(\omega)$ , meaning that the true state  $\omega$  belongs to the set of states that the agent considers possible, denoted by  $P(\omega)$ .

**2. Property 2 (Positive introspection):** If  $\omega' \in P(\omega)$  then  $P(\omega') \subseteq P(\omega)$ , meaning that if the agent considers  $\omega'$  possible at state  $\omega$ , then the agent's information set at  $\omega$  is a subset of the information set at  $\omega'$ .

**3. Property 3 (Negative Introspection):** If  $\omega' \in P$  then  $P(\omega') \supseteq P(\omega)$ , meaning that if the agent considers  $\omega'$  possible at state  $\omega$ , then the information set at  $\omega$  contains the information set at  $\omega'$ .

*Definition 7* A **partitional information structure** means that: The information structure is represented as a **partition** of the set  $\Omega$ , meaning that each state  $\omega \in \Omega$  belongs to exactly one element of the partition  $P(\omega)$ . The partition elements are **mutually exclusive** and **exhaustive** subsets of  $\Omega$ . Each partition element represents the set of states that are indistinguishable to the agent in that state.

If we combine properties 2 and 3 :  $P(\omega') \subseteq P(\omega), P(\omega') \supseteq P(\omega) \Rightarrow P(\omega') = P(\omega)$ . This means that for any two states  $\omega, \omega' \in \Omega$ , if  $\omega' \in P(\omega)$ , then  $P(\omega') = P(\omega)$ , implying that all states that are in the same information set are indistinguishable from each other. This satisfies the condition that the information sets form a **partition** of  $\Omega$ , where each state belongs to exactly

one partition element. Since the partition elements are mutually exclusive and exhaustive subsets of  $\Omega$ , we conclude that  $P$  is a partition. ■

Formally in Kripke logic:

In Kripke logic, the **truthfulness** property states that if agent  $i$  knows a proposition  $\varphi$ , then  $\varphi$  must hold in the actual world.

*Well-formed formula 27*

$$K_i\varphi \Rightarrow \varphi \Rightarrow \forall w \in W, \text{ if } w \models K_i\varphi \rightarrow w \models \varphi$$

In Kripke logic **positive introspection** can be presented as :

*Well-formed formula 28*

$$K_i\varphi \Rightarrow K_iK_i\varphi$$

This states that if agent  $i$  knows  $\varphi$ , then agent  $i$  also knows that they know  $\varphi$ .

In Kripke logic, **negative introspection** can be expressed as:

*Well-formed formula 29*

$$\neg K_i\varphi \Rightarrow K_i\neg K_i\varphi$$

This means that if agent  $i$  does not know  $\varphi$ , then agent  $i$  knows that they do not know  $\varphi$ . Common knowledge operator can be defined as follows:

*Well-formed formula 30*

$$CK(p) \Leftrightarrow \forall_i, K_i(p) \wedge \forall_i, K_i(KK(p)) \dots$$

This can also be modeled with fixed point logic<sup>10</sup> where common knowledge can be reached:

*Well-formed formula 31*

$$CK(p) \Leftrightarrow K_1(p) \wedge K_2(p) \wedge \dots \wedge K_n(p) \wedge K_1(KK(p)) \wedge K_2(KK(p)) \dots$$

### 3.5 Back to Kripke's S5 system

*Theorem 3* Kripke's S5 system:  $N, \Omega, \{T_i\} \in N$  is a knowledge space. And  $A \in 2^\Omega$  is an event. Where,  $N$  is a set of players,  $A$  is a finite set of actions,  $\Omega$  is the space of the states of the world. We will assume here that  $\Omega$  is finite.  $T_i$  is the space of possible types of player  $i$ . And,  $t_i: \Omega \rightarrow T_i$  is player  $i$ 's private signal or type. Information partitions is:  $P_i(\omega) = \{\omega': t_i(\omega') = t_i(\omega)\}$ , that is  $P_i(\omega)$  is the set of states of the world for which player  $i$  has the same type as he/she does in  $\omega$ . And,  $\omega_i \in P_i(\omega)$ , the set  $\{P_i(\omega)\}_{\omega \in \Omega}$  is easily seen to be a partition  $\Omega$ , and is called  $i$ 's information partition. A knowledge space can thus be given as:  $(N, \Omega, \{P_i\} \in N)$ , see [Tamuz \(2024\)](#).

<sup>10</sup> Consider the following iterative function representing mutual knowledge up to depth  $k$ :  $K^k(p) = K_1(K^{k-1}(p)) \wedge K_2(K^{k-1}(p))$ , where  $K^0(p) = p$ ;  $K^1(p) = K_1(p) \wedge K_2(p) \dots \wedge K_n(p)$ ,  $K^k(p)$ , for  $k > 1$  represents each agent knowing up to depth  $k$  that  $p$  holds. The process of iterating this mutual knowledge level  $K^k(p)$  will converge at a fixed point:  $CK(p) = \lim_{(k \rightarrow \infty)} K^k(p)$ .

1.  $K_i\Omega = \Omega$  A player knows that some state of the world has occurred. And given  $K_iA$  a set of states of world in which  $i$  knows  $A$  and  $A \in 2^\Omega$ :

equation 14

$$K_iA = \{\omega: P(\omega) \subseteq A\} \equiv K_iA = \bigcup \{\omega: P(\omega) \subseteq A\}$$

2.  $K_iA \cap K_iB = K_i(A \cap B)$ . A player knows  $A$  and a player knows  $B$  if and only if he knows  $A$  and  $B$ .
3. Axiom of knowledge:  $K_iA \subseteq A$  a player knows  $A$  then  $A$  has indeed occurred.
4. Axiom of positive introspection:  $K_iK_iA = K_iA$ . If a player knows  $A$  then he/she knows that he/she knows  $A$ .
5. Axiom of negative introspection:  $(K_iA)^c = K_i((K_iA)^c)$ . If a player does not know  $A$  then she knows that she does not know  $A$ .

*Proof:*

1. This follows from the definition
2.  $K_iA \cap K_iB = \{\omega: P_i(\omega) \subseteq A\} \cap \{\omega: P_i(\omega) \subseteq B\} = \{\omega: P_i(\omega) \subseteq A \cap B\} = K_i(A \cap B)$
3. If  $\omega \in K_iA$ , so that  $P_i(\omega) \subseteq A$ , since  $\omega_i \in P_i(\omega)$ , it follows that  $\omega \in A$  and so  $K_iA \subseteq A$ .
4. By the previous we have that  $K_iK_iA \subseteq K_iA$ . Now, let  $\omega \in K_iA$  so that  $P_i(\omega') = P_i(\omega)$ , so it follows that  $\omega' \in K_iA$ , and since  $\omega'$  is an arbitrary element of  $P_i(\omega)$  it was shown that  $P_i(\omega) \subseteq K_iA$ , and hence by definition  $\omega \in K_iK_iA$
5. The left-hand side  $(K_iA)^c$  represents the event that agent  $i$  does *not* know  $A$ . The right side,  $K_i((K_iA)^c)$  represents the event that agent  $i$  knows that they do not know  $A$ . In modal logic we apply positive introspection i.e. if an agent knows something, they know that they know it. Formally,  $K_iA \Rightarrow K_iK_iA$ . We also assume the negative introspection axiom i.e., if an agent does not know something, they know that they do not know it:  $(K_iA)^c = K_i((K_iA)^c)$  ■

#### 4. Reasoning Depth, Kantian equilibrium and Nash equilibrium

This literature on reasoning depth postulates that each player has a bound  $k$  on reasoning, where  $k \in \{0, 1, \dots\}$ . So, a player with  $k = 0$  is a nonrational and nonstrategic type which is allowed to take any action, and his behavior is used by other players to anchor their beliefs, see [Strzalecki, T. \(2014\)](#). But for a general Level –  $k$  reasoning we have, For any level  $k \geq 0$ , a Level- $k$  player maximizes their strategy by assuming that the other player is reasoning at Level  $k - 1$ . And mathematically for player  $P_1$  we have:

equation 15

$$S_1^k = \arg \max_{s_1 \in S_1} U_1(s_1, S_2^{k-1})$$

Where  $S_2^{k-1}$  is the strategy that  $P_2$  would select based on level  $K - 1$  reasoning. At  $S_1^0$  previous equals to:  $S_1^0 = \arg \max_{s_1 \in S_1} U_1(s_1, s_2^{default})$ , where  $s_2^{default}$  is some default assumption about  $P_2$  strategy<sup>11</sup>. Since the contribution by [Nagel \(1995\)](#), it is well established that limited depth of

<sup>11</sup> For level 1 reasoning we have:  $S_1^1 = \arg \max_{s_1 \in S_1} U_1(s_1, S_2^0)$ . A Level- 1 player, such as  $P_1$ , assumes that  $P_2$  is a Level-0 reasoner.

reasoning accounts for important features of experimental data which are missed by models of full rationality, see also [Cooper et al.\(2024\)](#). Here, we are not going to delve thoroughly into the literature on Kantian equilibrium but following [Osborne,M.J., Rubinstein,A.\(2023\)](#), and [Roemer \(2010\)](#), [Roemer \(2019\)](#), we will provide following definition:

**Definition 8:** A vector of strategies  $\mathcal{L} = (\mathcal{L}^1, \dots, \mathcal{L}^n)$  is a multiplicative Kantian equilibrium of the game  $G = S(V^1, \dots, V^n)$  for  $\forall i = 1, \dots, n$

equation 16

$$\arg_{\alpha \in \mathbb{R}_+} \max V^i(\alpha \mathcal{L}) = 1$$

Formally there is a set of  $n$  agents with payoff function  $V_i: \mathbb{R}_+^n \rightarrow \mathbb{R}$ , We define effort also as:  $\mathcal{L}^{-1} = (\mathcal{L}^1, \dots, \mathcal{L}^{i-1}, \mathcal{L}^{i+1}, \mathcal{L}^n)$ , and payoff function  $V_i$  is strictly monotone and decreasing in  $\mathcal{L}^{-1} \forall i$ .

**Definition 9:** In a strategic game  $\langle N, (A^i)_{i \in N} (\succsim^i)_{i \in N} \rangle$  an action profile  $a = (a^i) = A$  is a Nash equilibrium  $\forall i \in N : (a^i, a^{-i}) \succsim^i (x^i, a^{-i}), \forall x^i \in A^i$ . Where  $(x^i, a^{-i})$  denotes the action profile that differs from  $a$  only in that in action of individual  $i$  is  $x^i$  rather than  $a^i$ . And,  $N = (1, \dots, N)$  is a set of players, and  $u^i: A \rightarrow \mathbb{R}$  is a payoff function for player  $i$ . And for preferences,  $\forall i \in N \succsim^i$  over the set  $A^i = \times_{i \in N} A^i$  of action profiles.

**Definition 10:** Let  $\Gamma = \langle N, H, P(\succsim^i)_{i \in N} \rangle$  be an extensive game A strategy profile  $s$  is a Nash equilibrium of  $\Gamma$  if for every player  $i \in N$  we have:  $z(s) \succsim^i z(s^{-i}, r^i)$ , for every strategy  $r^i$  of player  $i$ . Where, for any strategy profile,  $\sigma(z(\sigma))$  is the terminal history generated by  $\sigma$ .

**Definition 11:** Let  $s$  be a strategy profile for the extensive game  $\langle N, H, P, (\succsim^i)_{i \in N} \rangle$ . The terminal history generated by  $s$  is  $(a_1, \dots, a_T)$  where  $a_1 = s^{P\emptyset}(\emptyset)$  and  $a_{t+1} = s^{P(a_1, \dots, a_t)}, t = 1, \dots, T - 1$ .

## 5. Bounded rationality: Luce model with bounded rationality included

Bounded rationality is understood as rationality exhibited by actual human economic behavior, see [Selten \(1998\)](#). Also see [Simon \(1957\)](#). Luce material draws from [Luce \(1959/2005\)](#).

**Definition 12 :**  $\rho$  has a Luce representation if there  $\exists w: X \rightarrow \mathbb{R}_{++}$ , where  $\rho$  is a stochastic choice function,  $X$  is a set of alternatives and:

equation 17

$$\rho(X, A) = \frac{w(X)}{\sum_{Y \in A} w(Y)}$$

Where  $A, B, C \subseteq X$  are finite choice problems or menus. A probability space here is:  $(\Omega, \mathcal{F}, \mathbb{P})$  and  $\mathcal{F}$  - measurable random utility function  $\tilde{U}: \Omega \rightarrow \mathbb{R}^X$ .

**Axiom 1** Let  $T \subseteq_{finite} U$  such that,  $\forall S \subset T, P_S$  is defined:

- If  $P(x, y) \neq 0, 1 \forall x, y \in T$ , then for  $R \subset S \subset T$   
 $P_T(R) = P_S(R)P_T(S)$ ;
- If  $P(x, y) = 0$  for some  $x, y \in T$ , then  $\forall S \subset T$

$$P_T(R) = P_{T-\{x\}}(S - \{x\});$$

*Axiom 2* The ordinary probability axioms are :

- For  $S \subset T$  ,  $0 \leq P_T(S) \leq 1$
- $P_T(T) = 1$
- If  $R, S \subset T$  and  $R \cap S = \varphi$ , then  $P_T(R \cup S) = P_T(R) + P_T(S)$

Independence of irrelevant alternatives, see [Luce, R. D. \(1977\)](#) :

*Lemma 3* if  $P(x, y) \neq 0, 1 \forall x, y \in T$ , then Axiom1 implies that  $\forall S \subset T$  such that  $x, y, \in S$

*equation 18*

$$\frac{P(x, y)}{P(y, x)} = \frac{P_S(x)}{P_S(y)}$$

Proof: By the Axiom 1 we know that:  $P_S(x) = P(x, y)[P_S(x) + P_S(y)]$

So now:

*equation 19*

$$P_S(x)[1 - P(x, y)] = P_S(x)P(y, x) = P(x, y)P_S(y) \blacksquare$$

One simple case of Luce model with bounded rationality will be shown in this section. With bounded rationality, the players are less sensitive to differences in utility, making the choice probabilities more balanced and less extreme than in a fully rational model. This dynamic can lead to slower convergence or even oscillations, depending on the balance between reinforcement and social influence. In this example parameters used are:  $\lambda = 0.5$ -this is bounded rationality parameter,  $\alpha = 0.1$  - Learning rate for reinforcement,  $\beta = 0.05$  - Social influence parameter,  $T = 100$ - Number of time steps.

*equation 20*

$$P(a_j) = \frac{e^{\lambda \cdot u(a_j)}}{\sum_{k=1}^n e^{\lambda \cdot u(a_k)}}$$

Where  $u(a_j)$  denotes utility of player choosing option  $a_j$  and the modified probability would be

:  $P_i(a_j) = \frac{u_i(a_j)}{\sum_{k=1}^n u_i(a_k)}$ . Dynamic version of previous equation include  $t$  superscript:  $P_i^t(a_j) =$

$\frac{u_i^t(a_j)}{\sum_{k=1}^n u_i^t(a_k)}$ . Now for reinforcement Learning: In reinforcement learning, utilities are updated based on past choices and outcomes. So, if player  $i$  chooses option  $a_j$  at time  $t - 1$  ,then:

*equation 21*

$$u_i^t(a_j) = u_i^{t-1}(a_j) + \alpha \cdot \delta_i^{t-1}$$

Where  $\alpha$  is learning rate and  $\delta_i^{t-1}$  represents the reward (or penalty) player  $i$  received from choosing  $a_j$  at time  $t - 1$ . Players may also update their utilities based on the choices of others, with a social influence factor. If player  $j$  has chosen option  $a_k$  more frequently, other players may increase their perceived utility of  $a_k$ :

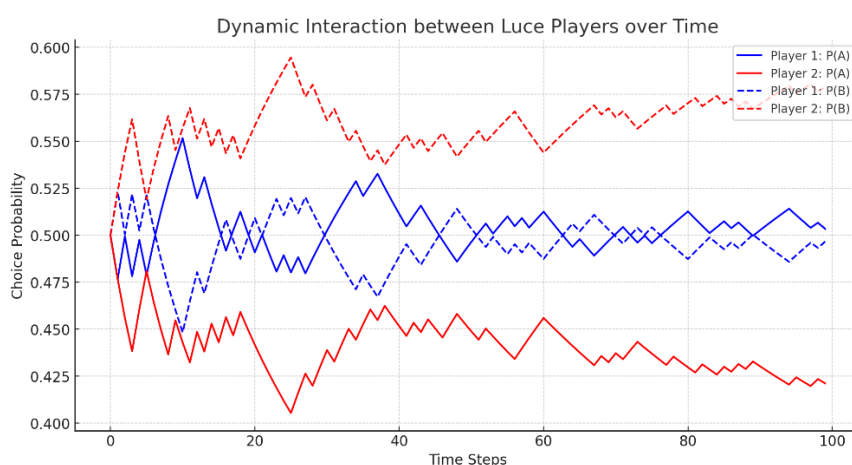


equation 22

$$u_i^t(a_k) = u_i^{t-1}(a_k) + \beta \sum_{j \neq i} P_j^{t-1}(a_k)$$

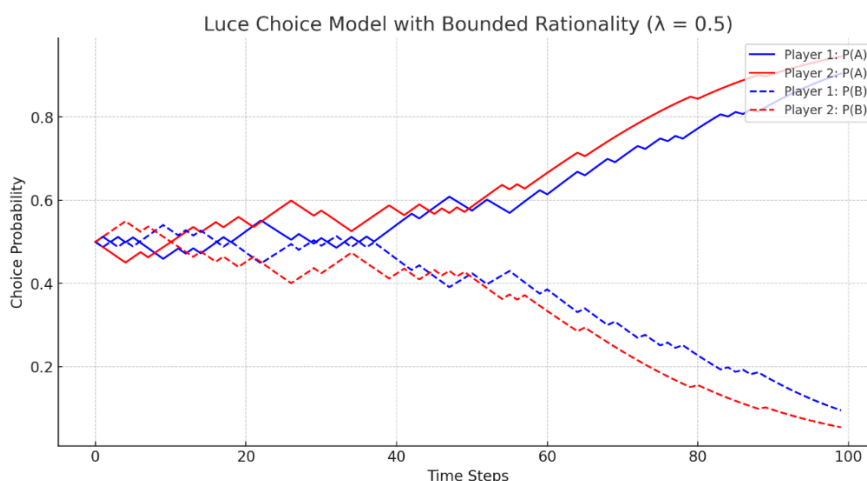
Where  $\beta$  is parameter governing the influence of others' choices on player  $i$ 's utility. Next, it will be presented dynamic market with competing brands. Imagine a market with two competing brands,  $A$  and  $B$ , and two players representing consumers who make probabilistic choices between the brands. Each consumer updates their perceived utility based on personal satisfaction (reinforcement) and by observing the choices of the other consumer (social influence).

Figure 1 Dynamic interaction between Luce players



Source: Author's own calculation

Figure 2 Luce's choice model with bounded rationality parameter  $\lambda = 0.5$



Source: Author's own calculation

### 5.1 Conditional probability theory and Luce model

The conditional probability of  $S$  given  $T$  such that  $p(T) > 0$  is defined as:

equation 23

$$p(S|T) = \frac{p(S \cap T)}{p(T)}$$

Now if  $R \subset S \subset T$ , following axiom 1 (If  $P(x, y) \neq 0, 1 \forall x, y \in T$ , then for  $R \subset S \subset T$ ;  $P_T(R) = P_S(R)P_T(S)$ ); then:

equation 24

$$p(R|S)p(S|T) = \frac{p(R \cap S)}{p(S)} \frac{p(S \cap T)}{p(T)} = \frac{p(R)}{p(S)} \frac{p(S)}{p(T)} = \frac{p(R \cap T)}{p(T)} = p(R|T)$$

For this section see more in [Rényi, A.\(1955\)](#). This is, of course, the formal analogue of part i of axiom 1. By taking three arbitrary sets, instead of  $R \subset S \subset T$ , a somewhat more general condition can be shown to hold.

## 5.2 Matching law formulation

Here it is formulated following theorem, and it is provided a proof.

*Theorem 1* Any matching law selection rule satisfies Luce's choice axiom. Conversely, if  $p(a|A) > 0 \forall a \in A \subset X$ , then Luce's choice axiom implies that it is a matching law selection rule.

*Proof:* The **matching law** states that the probability of selecting an option  $a$  from a set of alternatives  $A$  is proportional to some positive utility or "value"  $v(a)$  associated with  $a$ . Formally, a selection rule follows the matching law if:  $p(a|A) = \frac{v(a)}{\sum_{b \in A} v(b)}$ . Where  $v(a) > 0, \forall a \in A$ . Luce's choice axiom or Independence from Irrelevant Alternatives (IIA) states that for any two options  $a, b \in A$ :

equation 25

$$\frac{p(a|A)}{p(b|A)} = \frac{p(a|\{a, b\})}{p(b|\{a, b\})}$$

Under the matching law, we have:  $p(a|A) = \frac{v(a)}{\sum_{c \in A} v(c)}$ ;  $p(b|A) = \frac{v(b)}{\sum_{c \in A} v(c)}$  Hence:  $\frac{p(a|A)}{p(b|A)} =$

$\frac{\frac{v(a)}{\sum_{c \in A} v(c)}}{\frac{v(b)}{\sum_{c \in A} v(c)}}$ . Similarly for the set  $\{a, b\}$ :

equation 26

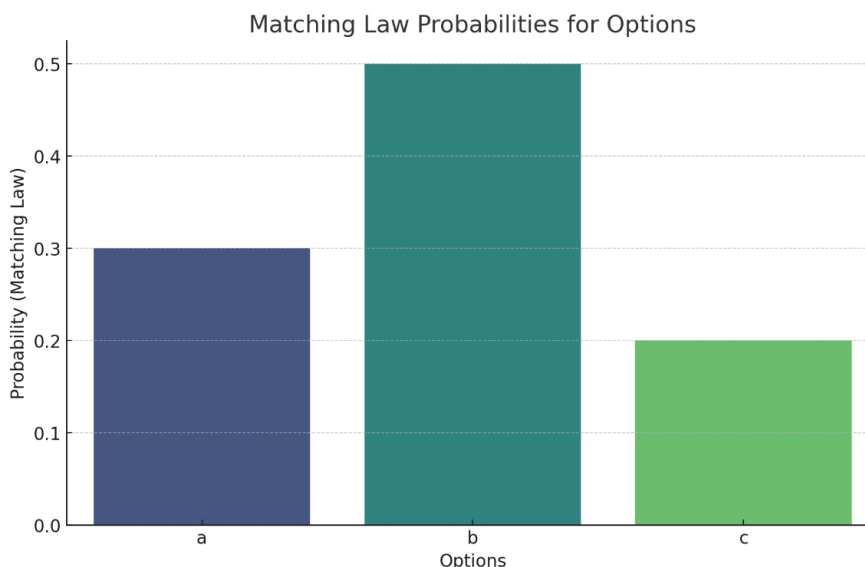
$$\frac{p(a|\{a, b\})}{p(b|\{a, b\})} = \frac{\frac{v(a)}{v(a) + v(b)}}{\frac{v(b)}{v(a) + v(b)}} = \frac{v(a)}{v(b)} \blacksquare$$

Since both ratios are equal, the matching law selection rule satisfies Luce's choice axiom. Next, we may code and plot this theorem. First, we will define utilities: we will assign positive arbitrary values  $v(a), \forall a \in A$ . Then we will calculate probabilities by the matching law:  $p(a|A) = \frac{v(a)}{\sum_{b \in A} v(b)}$ . And we will verify Luce's Choice Axiom: check for each pair of options  $a, b$  that  $\frac{p(a|A)}{p(b|A)} = \frac{v(a)}{v(b)}$ . The plot shows the matching law probabilities for each option based on their assigned utilities. Here's the summary of the ratios for each pair of options to verify Luce's choice axiom:

- Ratio  $a/b$ :  $\frac{p(a|A)}{p(b|A)} = 0.6; \frac{v(a)}{v(b)} = 0.6$
- Ratio  $a/c$ :  $\frac{p(a|A)}{p(c|A)} \approx 1.5; \frac{v(a)}{v(c)} = 1.5$
- Ratio  $b/c$ :  $\frac{p(b|A)}{p(c|A)} = 2.5; \frac{v(a)}{v(c)} = 2.5$

Since the probability ratios are equal to the utility ratios for each pair, the plot and calculations confirm that this selection rule satisfies Luce's choice axiom. The plot will be shown on the following page.

Figure 3 Matchin law probabilities for options



Source: Author's own calculation

## 6.The Agreement Theorem

[Aumann \(1976\)](#) posed the following question: could two individuals who share the same prior ever agree to disagree? See [Levin \(2016\)](#). That means if  $i, j$  share common previous beliefs over states of the world, could it be that state arise at which it was commonly known that  $i$

assigns probability of some event  $p_i$ , and  $j$  assigned probability of  $p_j$  and  $p_i \neq p_j$ . Aumann concluded that this sort of disagreement is impossible. Now, formally let  $p$  be a probability measure on  $\Omega$  which are agents' prior belief. For any state  $\omega$  and event  $E$ , let  $p(E|p_i(\omega))$  denote  $i$ 's posterior belief, so that  $p(E|p_i(\omega))$  is obtained under Bayes' rule. The event that agent  $i$  assigns probability  $p_i$  to  $E$  is  $\{\omega \in \Omega: p(E|p_i(\omega)) = p_i\}$

**Proposition 2** Suppose two agents have the same prior belief over a finite set of states  $\Omega$ . If each agent's information function is partitional and it is common knowledge in some state  $\omega \in \Omega$  that agent 1 assigns probability  $p_1$  to some event  $E$  and agent 2 assigns probability  $p_2$  to  $E$ , then  $p_1 = p_2$

*Proof:* If the assumptions are satisfied then there is some self-evident event  $F$  and  $\omega \in F$ :

equation 27

$$F \subset \{(\omega' \in \Omega: p(E|p_1(\omega')) = p_1) \cap \{(\omega' \in \Omega: p(E|p_2(\omega')) = p_2)\}\}$$

Since  $\Omega$  is finite, so is the number of sets in each union and let  $F = \cup_k A_k = \cup_k B_k$  and for a nonempty disjoint sets  $C, D$  with  $p(E|C) = p_i$  and  $p(E|D) = p_i$  we have that  $p(E|C \cup D) = p_i$ , and  $\forall k, p(E|A_k) = p_1$ , then  $p(E|F) = p_1$  and similarly  $p(E|F) = p(E|B_k) = p_2$  ■

## 7. Numerical examples

**First example:** Let's assume the initial beliefs of the three agents are as follows for the two worlds:

Table 1 agents initial beliefs

Agent 1: [0.6, 0.4]
Agent 2: [0.5, 0.5]
Agent 3: [0.7, 0.3]

In this python code defined parameters are

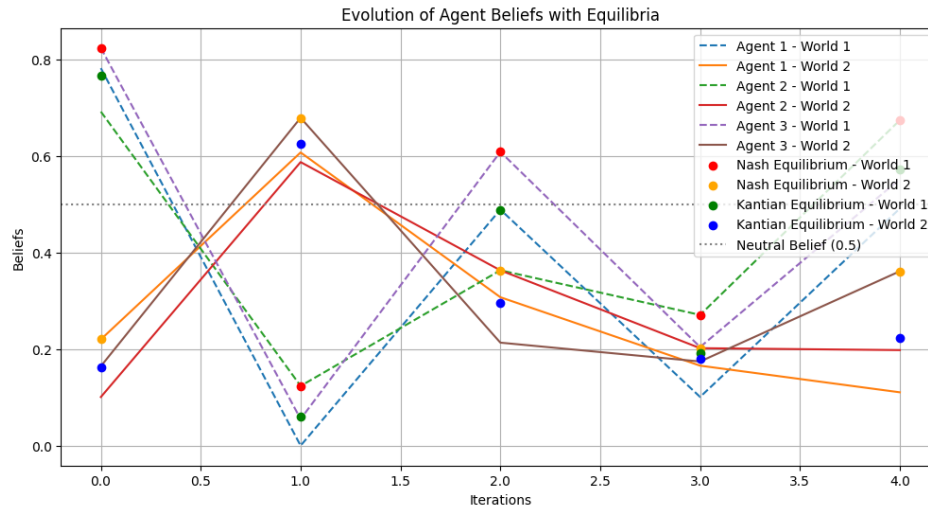
```
# Define parameters : n_agents = 3 # Number of agents , n_worlds = 2 # Number of worlds,
n_iterations = 5 # Number of iterations for belief updates, k_levels = [0, 1, 2] # Levels of cognitive
reasoning. We'll run the belief updates over 5 iterations to see how beliefs evolve. To simulate
k-level thinking, we can add some Gaussian noise to the beliefs during updates. For each
iteration, we'll calculate the Nash and Kantian equilibria based on the agents' beliefs.
```

Table 2

Iteration	Agent 1	Agent 2	Agent 3	Nash EQ	Kantian EQ
1	[0.78145853 0.21999514]	[0.69146899 0.10070053]	[0.82334435 0.16441218]	[0.82334435 0.21999514]	[0.76542396 0.16170262]
2	[0. 0.60697405]	[0.12356335 0.5870302]	[0.05509522 0.67869425]	[0.12356335 0.67869425]	[0.05955286 0.62423284]
3	[0.48865891 0.30727302]	[0.36306045 0.36296319]	[0.60827405 0.2132603]	[0.60827405 0.36296319]	[0.48666447 0.29449884]

	[0.10082487	[0.27067617	[0.20415334	[0.27067617	[0.19188479
4	0.16559914]	0.2018131]	0.17422747]	0.2018131]	0.18054657]
	[0.49093557	[0.67386333	[0.55193592	[0.67386333	[0.57224494
5	0.11045239]	0.19782752]	0.36106872]	0.36106872]	0.22311621]

Figure 4 Evolution of agents beliefs with Kantian and Nash equilibria

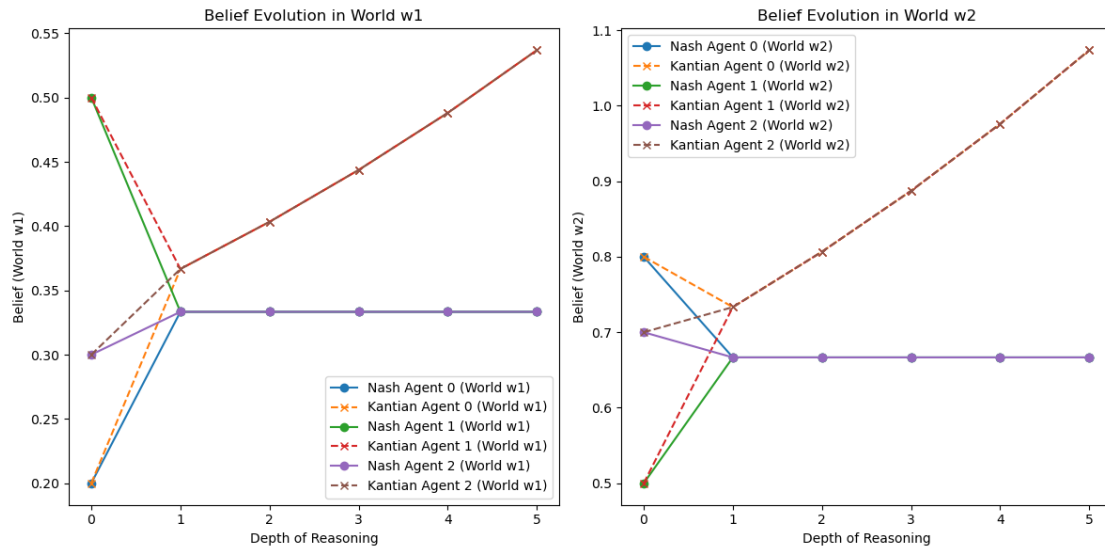


Source: Author's own calculation

This numerical example provides insight into how agents' beliefs evolve in a structured way and how those beliefs can converge to equilibria based on the reasoning level of the agents. This model captures the interactions between bounded rationality, beliefs, and the moral reasoning framework of Kantian equilibrium. Each agent's decisions are based on their cognitive depth while considering the moral implications of their actions and beliefs. This framework can be further elaborated into computational models or simulations to analyze the dynamics of belief convergence, the impact of different levels of reasoning, and the emergence of common knowledge under various settings.

**Second example:** To address this complex theoretical framework, let's integrate several core concepts into a single model by combining Kripke semantics, bounded rationality, common knowledge (especially relevant in Kripke semantics), Aumann's Agreement Theorem, Kantian and Nash equilibria, and k-level thinking in a cognitive hierarchy. I'll outline each element of the framework, then derive a model that applies this to an economic network that operates within a semantic economy. The goal is to showcase how agents interact in such a network to create informational value.

Figure 5



Source: Author's own calculation

Some equations of the model used for the previous plots are: Consider a Kripke model with a set of agents  $A = \{1, 2, \dots, N\}$  and two accessible worlds  $W = \{w1, w2\}$  representing different economic states.

**Agent Beliefs and Bounded Rationality:**

- Each agent  $i \in A$  holds beliefs over  $W$ , denoted as  $b_i(w)$  for world  $w$ .
- Agents update their beliefs iteratively with a bounded rationality constraint, applying  $k$ -level thinking up to their cognitive hierarchy level  $k_i$

**Belief Update Mechanisms:**

- **Nash Belief Update:** Agent  $i$ 's Nash belief update at level  $k_i$  depends on the mean beliefs of other agents in the network.

equation 28

$$b_i^{Nash}(w) = \frac{1}{N-1} \sum_{j \neq i} b_j(w)$$

- **Kantian Belief Update:** Each agent  $i$  updates beliefs by reflecting on the common benefit and adjusting towards a shared belief.

equation 29

$$b_i^{Kantian}(w) = b_i(w) + \alpha \left( \frac{1}{N} \sum_j b_j(w) \right)$$

Here,  $\alpha$  represents the level of introspection, capturing the Kantian principle of considering the joint outcome.

**Aumann's Agreement and Almost Common Knowledge:**

- After several rounds of belief updates, beliefs converge as per Aumann's Agreement Theorem when almost common knowledge is established—i.e., agents share a sufficiently high belief in each other's beliefs across worlds.

### Fixed-Point Condition in Cognitive Hierarchy:

- For agents to reach a **fixed-point** agreement, each agent's belief update (at any reasoning depth  $k_i$ ) should stabilize across worlds:

*equation 30*

$$b_i^*(w) = b_i(w), \forall i \in A; w \in W$$

### Semantic Economy and Informational Value Creation:

- In an economic network, agents share and evolve their beliefs to create informational value, not merely to reduce costs. This added value emerges as agents reach consensus and stabilize around shared insights, enhancing collective knowledge within the network.

**Third example:** To derive a mathematical model that unites **Kripke semantics**, **bounded rationality**, **common knowledge (CK)**, **Aumann's Agreement Theorem**, **Kantian Nash equilibrium**, **k-level reasoning**, and **Cognitive Hierarchy Theory** within the framework of an **economic network**, we must consider a few foundational elements. This model will illustrate how cognitive reasoning and network interactions influence economic advantage, not through information cost reduction, but by generating new informational value across a network.

### Components of the Model

1. **Kripke Semantics:** A framework for modeling knowledge and beliefs across possible worlds.
2. **Bounded Rationality:** Agents have limitations in processing power and information.
3. **Common Knowledge (CK):** Formally, an event  $ppp$  is common knowledge if everyone knows  $p$ , everyone knows that everyone knows  $p$ , ad infinitum. We will use fixed-point logic to represent this.
4. **Aumann's Agreement Theorem:** Two rational agents with common priors and knowledge of each other's beliefs cannot agree to disagree. They will reach a consensus if they continue to exchange information.
5. **Kantian Nash Equilibrium (KNE):** Each agent chooses a strategy that they would want all agents to adopt collectively.
6. **K-level Thinking in Cognitive Hierarchy Theory:** Each agent is modeled as having a finite depth of reasoning. Agent 0 acts without considering others, agent 1 considers agent 0's reasoning, and so forth.

7. **Networked Economic Context:** Agents interact in an economic network where competitive advantage depends on the informational value generated across the network.

## Model Assumptions and Definitions

- **Agents**  $A = \{1, 2, \dots, N\}$  in a networked economy.
- **Possible Worlds:** Each agent has beliefs over a set of possible worlds  $W$ .
- **Information Sets:** Each agent  $i$  has a finite set of beliefs  $B_i$  about events and possible worlds, representing bounded rationality.
- **Common Knowledge (CK):** Defined through a fixed-point condition across agent beliefs in a given world  $w$ .
- **Kantian Nash Equilibrium (KNE):** Agents choose strategies for a common good, reflecting rational expectations across the network.
- **K-level Reasoning:** Each agent reasons up to level  $k_i$  varying by agent. This represents the cognitive hierarchy in the network.

### Step 1: Define Knowledge in Terms of Kripke Semantics

- Let  $W$  be the set of possible worlds, with  $w \in W$  representing a particular state of the world.

Each agent  $i$  has:

- **Access Relation  $R_i$ :** An agent's relation on  $W$  that reflects their perspective (or information) about the different possible worlds.
- **Knowledge representation:**  $K_i(p)$ : Where  $K_i(p) = \{w \in W \mid p \text{ is true in all accessible worlds by } R_i\}$
- **For common knowledge (CK) of event  $p$ ,** it holds that:  
 $CK(P) \Leftrightarrow K_1(p) \wedge K_2(p) \dots \dots \wedge K_N(p) \wedge K_1(KK(p)) \wedge K_2(KK(p)) \dots$

This means that for  $p$  to be common knowledge, each agent must know  $p$  and know that others know  $p$ , iteratively, which creates a fixed-point in logic.

### Step 2: Represent Bounded Rationality and K-Level Thinking

Each agent  $i$  can reason up to a finite **depth**  $k_i$ , which represents bounded rationality within **Cognitive Hierarchy Theory**:

- **Agent 0** has no reasoning about others.
- **Agent  $k$**  believes that others reason up to depth  $k - 1$

Let  $B_i^k$  represent the beliefs of agent  $i$  at reasoning depth  $k$ . If  $k \rightarrow \infty$  agents would theoretically reach full common knowledge; however, bounded rationality limits  $k_i$ .

### Step 3: Incorporate Aumann's Agreement Theorem and Kantian Nash Equilibrium

1. **Aumann's Agreement Theorem** implies that if agents  $i$  and  $j$  have common priors and share beliefs up to common knowledge, they cannot "agree to disagree."



2. For Kantian Nash Equilibrium (KNE), agents maximize **social utility** by adopting strategies they would prefer all agents to adopt, which is consistent with cooperative decision-making.

Define each agent's **utility** function in the economic network as:

*equation 31*

$$U_i(s) = f(s) + g(s, B_i^k) + h(B_j^{k-1})$$

$f(s)$ -represents strategy dependent payoff,  $g(s, B_i^k)$  -personal information dependent component,  $h(B_j^{k-1})$ :Network (others' beliefs)-dependent component, capturing information value.

#### **Step 4: Competitive Advantage through Informational Value**

In a **semantic economy**, agents' competitive advantage depends on the informational value they create for the network:

*equation 32*

$$V_i = \sum_{j \in A} h(B_j^{k-1}) - cost(B_i)$$

Where here:

- **informational Value**  $V_i$ : Derived from the sum of other agents' knowledge contributions.
- **Cost**: The cost of acquiring or processing information.

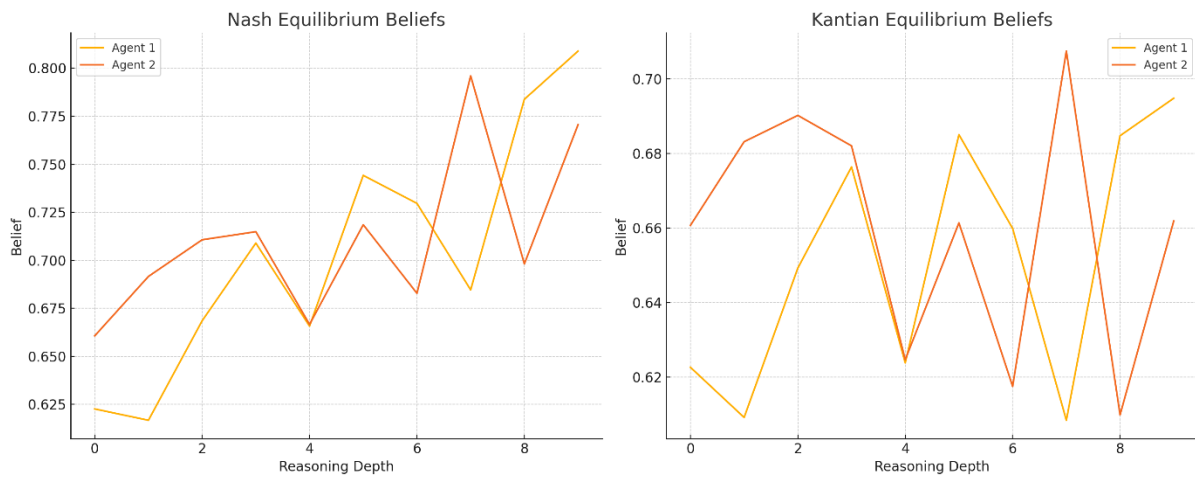
Thus, an agent maximizes utility by maximizing  $V_i$  rather than minimizing information acquisition costs. This value-driven approach leads agents to choose strategies that enhance collective knowledge.

Parameters in following simulation are:

$n_{agents} = 2$  # Number of agents  $n_{depths} = 10$  # Levels of reasoning depth  $true_{probability} = 0.65$   
# Initial true probability in the world. In this model and simulation model is improved to ensure:

1. Depth of reasoning variation in the belief values for each equilibrium, showing progressive adjustments as reasoning depth increases.
2. Distinct Nash and Kantian adjustments reflect the Kantian emphasis on aligned beliefs and mutual benefit, while the Nash framework remains based on best responses to others' beliefs at each depth level.

Figure 6



Source: Author's own calculation

The plots show the belief evolution across reasoning depths, highlighting the differences between Nash and Kantian equilibria.

Table 3 Nash Equilibrium Beliefs

Agent	Depth 0	Depth 1	Depth 2	Depth 3	Depth 4	Depth 5	Depth 6	Depth 7	Depth 8	Depth 9
Agent 1	0.6226	0.6167	0.6685	0.7090	0.6657	0.7442	0.7296	0.6846	0.7838	0.8089
Agent 2	0.6607	0.6916	0.7107	0.7149	0.6666	0.7185	0.6828	0.7960	0.6981	0.7706

Source: Author's own calculation

Table 4 Kantian Equilibrium Beliefs

Agent	Depth 0	Depth 1	Depth 2	Depth 3	Depth 4	Depth 5	Depth 6	Depth 7	Depth 8	Depth 9
Agent 1	0.6226	0.6092	0.6492	0.6763	0.6238	0.6850	0.6598	0.6084	0.6847	0.6948
Agent 2	0.6607	0.6831	0.6902	0.6820	0.6246	0.6614	0.6175	0.7074	0.6098	0.6619

Source: Author's own calculation

Nash versus Kantian Equilibrium: The Nash equilibrium beliefs show a slightly greater upward trend as the depth of reasoning increases. This is because Nash beliefs adjust based on others' best responses, while Kantian beliefs account for mutual benefit, dampening large shifts.

## 8. Conclusion

Simulation of Kripke semantics, bounded rationality, common knowledge, Aumann's agreement theorem, Kantian Nash equilibrium, and k-level thinking within the framework of Cognitive Hierarchy Theory (CHT), captures the interactions between bounded rationality, beliefs, and the moral reasoning framework of Kantian equilibrium. Each agent's decisions are based on their cognitive depth while considering the moral implications of their actions and beliefs. When we estimate deviations of beliefs, we have in mind that: Nash equilibrium are the beliefs that maximize each agent's utility given the others' beliefs. Kantian Equilibrium

represents the average beliefs of all agents at a specific iteration, reflecting a consensus view. Results from the first example show that agents change their beliefs from the initial values. Simulations are done for 3 agents in 2 worlds, and evolution of beliefs is shown in 5 iterations. K-level thinking with Gaussian noise was employed here too. Results show that agents 1,2,3 beliefs differ from Kantian and Nash equilibria. At 50% of iterations Kantian equilibrium is identical to neutral belief. Nash equilibrium follows Kantian equilibrium in the two worlds but exerts larger shifts from initial beliefs. At  $\frac{3}{4}$  of iterations agents 1,2,3 beliefs converge with Kantian and Nash equilibria, and they diverge once again. This is in line with the idea that small departures from common knowledge can have a dramatic effect on the set of equilibria. So, even if each layer is certain about the payoffs structure, even small incremental uncertainty about other's information can eliminate equilibria that exists when payoffs are common knowledge. This naturally leads to further explorations on this topic and some striking examples in the literature such as [Rubinstein \(1989\)](#). The fact that small perturbations of the information structure can eliminate Nash equilibria occurs because the Nash equilibrium correspondence (mapping from the parameters of the game to the set of equilibrium strategies) is not lower semi-continuous, see [Levin \(2016\)](#). The second example shows belief Evolution in world w1 and w2: The plots show each agent's belief trajectory over different reasoning depths. Nash equilibrium beliefs tend to cluster closely, while Kantian beliefs evolve with greater alignment, showing the cooperative tilt in introspective (Kantian) reasoning. Nash beliefs: Each agent averages the beliefs of all other agents. Over several reasoning layers, these beliefs evolve toward a shared understanding, though agents are individually optimizing. When the reasoning depth increases, players recognize more intricate patterns in others' intentions and responses. This depth often leads to more robust cooperative expectations, as players anticipate others' willingness to adopt strategies that align more closely with a collective or "Kantian" principle. As a result, the Kantian beliefs about what is rational or optimal under cooperative strategies become stronger or "shift up." Nash equilibria differ from initial true probability in the world

## References

1. Arrow, K. J.; Debreu, G. (1954). Existence of an equilibrium for a competitive economy. *Econometrica*. **22** (3): 265–290.
2. Aumann, R. J. (1976). Agreeing to Disagree. *The Annals of Statistics*, 4(6), 1236–1239. <http://www.jstor.org/stable/2958591>
3. Cooper, D., Fatas, E., Morales, A. , Qi, S. (2024). Consistent Depth of Reasoning in Level-k Models. *American Economic Journal: Microeconomics*. 16. 40-76. 10.1257/mic.20210237.

4. Dickson, M. (2007). Non-relativistic quantum mechanics, In Handbook of the Philosophy of Science, Philosophy of Physics, North-Holland, Pages 275-415,
5. Fourny, G. (2018). Kripke Semantics of the Perfectly Transparent Equilibrium. 10.3929/ethz-b-000277118.
6. Geanakoplos, J. (1989/2021). Game Theory Without Partitions, and Applications to Speculation and Consensus. The B.E. Journal of Theoretical Economics, vol. 21, no. 2, 2021, pp. 361-394. <https://doi.org/10.1515/bejte-2019-0010>
7. Geanakoplos, J. (1992). Common Knowledge. The Journal of Economic Perspectives, 6(4), 53–82. <http://www.jstor.org/stable/2138269>
8. Harsanyi, J. C. (1967). Games with Incomplete Information Played by “Bayesian” Players, I-III. Part I. The Basic Model. Management Science, Vol 14 No 3 November, 1967
9. Harsanyi, J. C. (1968). Games with Incomplete Information Played by “Bayesian” Players, I-III. Part II. Bayesian Equilibrium Points. Management Science, 14(5), 320–334. <http://www.jstor.org/stable/2628673>
10. Harsanyi, J. C. (1968). Games with Incomplete Information Played by “Bayesian” Players, I-III. Part III. The Basic Probability Distribution of the Game. Management Science, 14(7), 486–502. <http://www.jstor.org/stable/2628894>
11. Hintikka, J. (1962). Knowledge and Belief. Ithaca, N.Y.: Cornell University Press
12. Ilik, D., Lee, G., Herbelin, H. (2010). Kripke Models for Classical Logic, Annals of Pure and Applied Logic, Volume 161, Issue 11, Pages 1367-1378.
13. Jech, T. (1997) Set Theory, 2nd ed. New York: Springer-Verlag.
14. Kreps, D. (1977). A Note on Fulfilled Expectations Equilibria, J. Econ. Theory.
15. Kripke, S.A. (1959). A completeness theorem in modal logic" J. Symbolic Logic , 24 (1959) pp. 1–14
16. Kripke, S.A. (1962). The undecidability of monadic modal quantification theory. Z. Math. Logik Grundl. Math., 8, pp. 113–116
17. Kripke, S.A. (1963), Semantical analysis of modal logic, I, Z. Math. Logik Grundl. Math., 9 pp. 67–96
18. Kripke, S.A. (1965). Semantical analysis of modal logic, II. J.W. Addison (ed.) L. Henkin (ed.) A. Tarski (ed.), The theory of models, North-Holland pp. 206–22.
19. Levin, J. (2016). Knowledge and Equilibrium. Online lecture notes at: <https://web.stanford.edu/~jtlevin/Econ%20286/Knowledge%20and%20Equilibrium.pdf>
20. Luce, R. D. (1959/2005) Individual Choice Behavior: A Theoretical Analysis. New York: Wiley. Reprinted by Dover Publications.
21. Luce, R. D. (1977). The choice axiom after twenty years. Journal of Mathematical Psychology, 15(3), 215–233.
22. McKenzie, Lionel W. (1954). On Equilibrium in Graham's Model of World Trade and Other Competitive Systems. Econometrica. 22 (2): 147–161.
23. Milgrom, P. and N. Stokey (1982). Information, Trade and Common Knowledge, J. Econ. Theory.
24. Myerson, R. (1991), Game Theory: The Analysis of Conflict, Harvard University Press, Cambridge, MA
25. Nagel, R. (1995). Unraveling in Guessing Games: An Experimental Study. The American Economic Review, 85(5), 1313–1326. <http://www.jstor.org/stable/2950991>
26. Osborne, M.J., Rubinstein, A. (2023). Models in Microeconomic Theory, (Expanded Second Edition), Cambridge, UK: Open Book Publishers
27. Rényi, A. (1955). On a new axiomatic theory of probability, Acta Mathematica, 6, pp. 285–335
28. Roemer, J. E. (2010). Kantian Equilibrium. Scandinavian Journal of Economics, 112(1), 1–24.

29. Roemer, J. E. (2019). *How We Cooperate: A Theory of Kantian Optimization*. Yale University Press
30. Rubinstein, A. (1989). The Electronic Mail Game: Strategic Behavior Under "Almost Common Knowledge. *The American Economic Review*, 79(3), 385–391. <http://www.jstor.org/stable/1806851>
31. Rubinstein, A. (1998). *Modeling Bounded Rationality*, The MIT Press
32. Rubinstein, A. (2021). *Modeling Bounded Rationality in Economic Theory: Four Examples* in *Routledge Handbook of Bounded Rationality*, edited by Riccardo Viale, Routledge, 423-435. pdf
33. Selten, R. (1998). Features of experimentally observed bounded rationality. *European Economic Review*, 42(3-5), 413–436. doi:10.1016/s0014-2921(97)00148-7
34. Simon, H.A. (1957). *Models of Man: Social and Rational*. Wiley, New York.
35. Stalnaker R (1994) On the evaluation of solution concepts. *Theory and Decision* 37(1):49–73, DOI 10.1007/BF01079205, URL <https://doi.org/10.1007/BF01079205>
36. Stone, Jon R. (1996). *Latin for the Illiterati: Exorcizing the Ghosts of a Dead Language*. London: Routledge.
37. Strzalecki, T. (2014). Depth of reasoning and higher order beliefs, *Journal of Economic Behavior & Organization*, Volume 108, Pages 108-122,
38. Tamuz, O. (2024). Undergraduate game theory lecture notes. California Institute of Technology caltech
39. Tirole, J. (1982). On the Possibility of Speculation under Rational Expectations, *Econometrica*.
40. Zermelo, E. (1930). Über Grenzzahlen und Mengenbereiche. *Fund. Math.* 16, pp. 29-47.