A COMPUTATIONAL LEARNING SEMANTICS FOR INDUCTIVE EMPIRICAL KNOWLEDGE

KEVIN T. KELLY

Abstract. This paper presents a new semantics for inductive empirical knowledge. The epistemic agent is represented concretely as a learner who processes new inputs through time and who forms new beliefs from those inputs by means of a concrete, computable learning program. The agent's belief state is represented hyper-intensionally as a set of time-indexed sentences. Knowledge is interpreted as avoidance of error in the limit and as having converged to true belief from the present time onward. Familiar topics are re-examined within the semantics, such as inductive skepticism, the logic of discovery, Duhem's problem, the articulation of theories by auxiliary hypotheses, the role of serendipity in scientific knowledge, Fitch's paradox, deductive closure of knowability, whether one can know inductively that one knows inductively, whether one can know inductively that one does not know inductively, and whether expert instruction can spread common inductive knowledge—as opposed to mere, true belief—through a community of gullible pupils.

Keywords: learning, inductive skepticism, deductive closure, knowability, epistemic logic, serendipity, inference, truth conduciveness, bounded rationality, common knowledge.

1. Introduction

Science formulates general theories. Can such theories count as knowledge, or are they doomed to the status of *mere* theories, as the anti-scientific fringe perennially urges? The ancient argument for inductive skepticism urges the latter view: no finite sequence of observations can rule out the possibility of future surprises, so universal laws and theories are unknowable.

A familiar strategy for responding to skeptical arguments is to rule out skeptical possibilities as "irrelevant" (Dretske 1981). One implementation of that strategy, motivated by possible worlds semantics for subjunctive conditionals, is to ignore worlds "distant from" or "dissimilar to" the actual world. If you are really looking at a cat on a mat under normal circumstances, you wouldn't be a brain in a vat hallucinating a non-existent cat if it weren't there, so your belief is *sensitive* to the truth (Nozick 1981, Roush 2007). Or if you were to believe that there is a cat on the mat, most worlds in which your belief is false are remote worlds involving systematic hallucinations, so your belief is *safe* (Sosa 1999, Pritchard 2007, Williamson 2000).

So much for "ultimate", brain-in-a-vat skepticism applied to particular perceptual beliefs. But what about inductive skepticism concerning general scientific laws and theories? Belief in such laws and theories does not seem "safe". For example, if the true law were not of the form $Y = bX + a$, would science have noticed *already*? Are all worlds in which the true law has form $Y = cX^2 + bX + a$ safely bounded away from $Y = bX + a$

worlds in terms of similarity—regardless how small c is?[1] One can formally ignore small values of c by the *ad hoc* assumption that they are farther from the $c = 0$ world than are worlds in which c is arbitrarily close to 0. But the resulting discontinuity in similarity is questionable and, in any event, the conditional "if there were a quadratic effect, it would have been so large that we would have noticed it already" is implausible, however one contrives to satisfy it. In fact, the history of science teaches that we have been wrong on fundamental matters in the past, due to pivotal but small effects (e.g., the relativistic corrections to classical mechanics), and that we cannot guard against more such surprises in the future (Laudan 1981). So although subjunctive semantics appears to provide a plausible response to ultimate, brain-in-a-vat skepticism concerning ordinary perceptual knowledge, it is still overwhelmed by inductive skepticism, since, in that case, the nearby possibilities are exactly the *skeptical* ones.

The best that one can expect of even ideally diligent, ongoing scientific inquiry is that it detect and root out error *eventually.* So if there is inductive knowledge, it must allow for a time lag between the onset of knowledge and the detection and elimination of error in other possible worlds. There is a venerable tradition, expounded by Peirce (1878), James (1898), Reichenbach (1949), Carnap (1945), Putnam (1963), and Gold (1967) and subsequently developed by computer scientists and cognitive scientists into a body of work known as computational learning theory (Jain et al. 1999), that models the epistemic agent as a *learner* who processes information through time and who stabilizes, eventually, to true, inductive beliefs.

Inductive learning is a matter of finding the truth eventually. It is natural to think of inductive knowledge that ϕ as having learned that ϕ. Having learned that ϕ implies that one has actually stabilized to true belief that ϕ and that one would have *converged to true belief* whether ϕ otherwise. The proposed semantics is more lenient—one has knowledge that ϕ if and only if one has actually converged to true belief that ϕ (as in having learned) and one would have *avoided error* whether ϕ otherwise—one might simply suspend belief forever if the data are so unexpected that one no longer knows what is going on. Allowance for suspension of belief agrees better with scientific practice. Moreover, it turns out to be necessary if the consequences of known theories are to be knowable by the same standard.[2]

The semantics is not proposed as a true analysis of inductive knowledge in the traditional, exacting sense.[3] There may be no such thing, and it may not matter whether there is, since what matters in philosophy is not so much how we do talk, but how we *will* talk, after shopping in the marketplace of ideas. In that spirit, the semantics is proposed as a useful, unified, explanatory framework for framing prolems and conceptual issues at the intersection of inductive knowledge, inductive learning, information, belief, and time.

[1]Nozick (1981) and Roush (2007) argue that we would have noticed the failure of known laws already because, if a given uniformity weren't true, some distinct uniformity would have been. But in the polynomial example, all the regularities are law-like. Nor can one object that all linear laws are closer to a linear law than any quadratic law is, since the knowledge claim in question is that the true law is linear, so sensitivity forces one to move to non-linear laws. Vogel (1987) presents additional objections to tracking as an adequate account of inductive knowledge.

[2]Alternatively, one could simply stipulate that the deductive consequences of inductive knowledge are known (Roush 2007), but then one would have no explanation why or how they are known, aside from the stipulation.

[3]An long list of improvements is provided just prior to the conclusion.

Such issues include: the relation of learnability to knowability, how deductive inference produces new inductive knowledge from old, how inductive knowledge can thrive in a morass of inconsistency, why scientific knowledge should allow for a certain kind of luck or "serendipity", how one can know that one knows, why one can't know that one doesn't know, how to know your own Moore sentence, and how expert instruction can spread common inductive knowledge through a population of passive pupils.

One common thread running through the the development that follows is *epistemic parasitism*. Inference is not an argument or a mere, formal relation. It is the execution of a procedure for generating new beliefs from old. If inference produces new knowledge from old, it is because the inference procedure is guaranteed to produce new beliefs that satisfy the truth conditions for knowledge from beliefs that already do. Therefore, the semantics *explains how*, rather than merely *assumes that*, certain patterns of inference turn old knowledge into new. The basic idea is that the new knowledge is *parasitic* on the old because the inference pattern generates beliefs whose convergence tracks the convergence of the given beliefs. A related theme is the *hyper-intensionality* of belief. It is not assumed that the belief state of the agent is deductively closed or consistent, or that the learning method of the agent follows some popular conception of idealized rationality. Rather, rationality is something a computable agent can only approximate, and the desirability of doing so should be explained, rather than presupposed, by the semantics of learning and knowledge.

Inclusion of the entire learning process within models of epistemic logic is consonant with the current trend in epistemic logic (van Benthem 2011) toward more dynamic and potentially explanatory modeling of the agent. Recently, there have been explicit studies of truth tracking and safety analyses of knowledge (Holliday 2013) and of inductive learning within a modal logical framework (Gierasimczuk 2010). Earlier, Hendricks (2001) proposed to develop learning models for inductive knowledge, itself, and a rather different proposal was sketched, informally, in (Kelly 2001).[4] This paper places the latter approach on a firm, formal basis. For decades, Johan van Benthem has strongly encouraged the development of connections between learning theory and epistemic logic, both personally and in print, so it is a particular pleasure to contribute this study to his festschrift.

2. Syntax

Let $G = \{1, \ldots, N\}$ be indices for a group of N individuals. Let $\mathbf{L}_{\mathsf{atom}} = \{\mathsf{p}_i : i \in \mathbb{N}\}$ be a countable collection of atomic sentences. Define the modal language $\mathbf{L}_{\mathsf{BIT}}$ (belief, information, time) in the usual way, with the classical connectives, including $\perp$, and the modal operators presented in the following table, where Δ is understood to be a finite subset of $\mathbf{L}_{\mathsf{BIT}}$. The unusually rich base language reflects Scott's (1970) advice to seek more interesting epistemic principles in interactions among operators. In the following glosses, let t^* be the time of the *epistemic context* at which "i knows that ϕ" is assessed and let $t \geq t^*$ be the *time of evaluation*, which may lie in the future, due to the evaluation of a future tense operator. The aim is to analyze convergent belief that ϕ *was* true at t^*, so one must keep a "clean copy" of t^* in the model in order to determine whether i

[4]The differences are described, in detail, below.

believes at some later time $t > t^*$ that ϕ was true at t^* (Kamp 1971).[5]

Time

$@_t\,\phi$	*At:* it is true at t that ϕ.
$\mathsf{N}\,\phi$	*Now:* It is true at t^* that ϕ.
$\langle\mathsf{F}\rangle\,\phi$	*Future tense:* it is true at $t' \geq t$ that ϕ.
$\langle\dot{\mathsf{F}}\rangle\,\phi$	*Future context tense:* In epistemic context $t^{**} \geq t^*$, it is true that ϕ.

Information and Belief

$[\,\mathsf{I}\,]_i\,\phi$	*Information:* information has been made available to i by t^* that ϕ is true at t^*.
$[\mathsf{D}]_i\,\phi$	*Determination:* it is determined by information available to i at t^* and by the method of i at t^* that ϕ is true at t^*.
$[\mathsf{B}]_i\,\phi$	*Virtual belief:* the learning method of i at t^* directs i to believe that ϕ is true at t^*.

Methodology

$\langle\mathsf{M}\rangle_i\,\phi$	*Methodological feasibility:* it is feasible for i that ϕ is true.
$\psi\ \langle\mathsf{MD}]\!\!\rightarrow_{i,\Delta}\ \phi$	*Conditional methodological feasibility:* given that ψ is true, it is feasible for i to ensure that ϕ is true without altering i's learning disposition concerning the truth of the premises in Δ.
$\mathsf{S}_i\,\Delta$	*Inferential stability*: if i modifies her method in a way that holds her learning disposition with respect to statements in Δ fixed, then her future beliefs concerning the statements in Δ also remain unaltered—because i is insensitive to any changes in her sensory inputs that might result when other agents notice the changes to her method.

Let $\mathbf{L}_{@\mathsf{BIT}}$ denote the set of all $\mathbf{L}_{\mathsf{BIT}}$ sentences that are prefixed by an operator $@_t$ for some $t \in \mathbb{N}$. Extend $\mathbf{L}_{\mathsf{BIT}}$ with definitions as follows. For primitive operators $[\mathsf{X}]$, $\langle\mathsf{Y}\rangle$, introduce the dual operators:

$$\langle\mathsf{X}\rangle\,\phi := \neg[\mathsf{X}]\neg\,\phi;\quad [\mathsf{Y}]\,\phi := \neg\langle\mathsf{Y}\rangle\neg\phi.$$

A tilde above a box operator $[\mathsf{X}]_i$ indicates the "whether" form of the operator, which is defined as follows unless noted otherwise:

$$[\tilde{\mathsf{X}}]\,\phi \quad := \quad [\mathsf{X}]\,\phi\ \vee\ [\mathsf{X}]\neg\phi.$$

Introduce the standard notation:

$$\mathsf{B}_i := [\mathsf{B}]_i;\quad \mathsf{F} := \langle\mathsf{F}\rangle;\quad \mathsf{G} := [\mathsf{F}];$$

and similarly for $\dot{\mathsf{F}}$, $\dot{\mathsf{G}}$. Clean up notation in the following way:

$$\begin{aligned} \mathsf{S}_i\,\delta &\ :=\ \mathsf{S}_i\{\delta\}; \\ \psi\,\langle\mathsf{MD}]\!\!\rightarrow_{i,\delta}\ \phi &\ :=\ \psi\,\langle\mathsf{MD}]\!\!\rightarrow_{i,\{\delta\}}\ \phi. \end{aligned}$$

[5]Alternatively, one could introduce first-order quantifiers over temporal variables, but it is conceptually vivid to treat tense as a modality freely permutable with other modalities.

When Γ, Δ are finite subsets of $\mathbf{L}_{\mathsf{BIT}}$ and X_i is an arbitrary modal operator, let:

$$\begin{aligned} \mathsf{X}_i\Gamma &:= \bigwedge_{\gamma\in\Gamma} \mathsf{X}_i\,\gamma; \\ \Delta \to \Gamma &:= \bigwedge_{\delta\in\Delta} \delta \to \bigwedge_{\gamma\in\Gamma} \gamma; \end{aligned}$$

3. Computational Learning Models

Let E denote the set of possible *external worlds*. In a Kantian spirit, learning semantics imposes no structure or restrictions whatever on E. Let $T = \mathbb{N}$ be interpreted as discrete *stages of inquiry*. Let $G = \{1, \ldots, N\}$ be interpreted as a finite set of *agents*. Agent $i \in G$ is assumed to have some overall, discrete, physical sensory state at t that will be called the agent's current *input* at t. Think of $S = \mathbb{N}$ as code numbers for possible inputs. Inputs are not assumed to have propositional meanings (they are never assigned truth values), but their occurrence makes propositional information *available*. Let S^* be the set of all finite sequences of inputs, so each $\sigma \in S^*$ is a possible *input history*.

It is assumed that each agent's belief state is maintained by a *learning function* L that returns a verdict (1 for "believe" and 0 for "don't believe") for each sentence ϕ in $\mathbf{L}_{@\mathsf{BIT}}$ in light of the current input history σ:

$$L : S^* \times \mathbf{L}_{@\mathsf{BIT}} \to \{0, 1\}.$$

Let $\phi_c(x, y)$ be the binary partial recursive function computed by the Turing machine with Gödel index c.[6] Learning function L is computable if and only if there exists $c \in \mathbb{N}$ such that:

$$L_c(\sigma, \phi) \;=\; \phi_c(\langle\sigma\rangle, \ulcorner\phi\urcorner),$$

for all $\sigma \in S^*$ and $\phi \in \mathbf{L}_{@\mathsf{BIT}}$, where $\langle.\rangle$ is an effective encoding of S^* and $\ulcorner.\urcorner$ is an effective Gödel numbering of $\mathbf{L}_{@\mathsf{BIT}}$. Let C denote the set of all $c \in \mathbb{N}$ such that $\phi_c(\langle.\rangle, \ulcorner.\urcorner))$ is a learning function. Elements of C are called *learning methods*.

Each learning method covers all future contingencies, but i's learning method can change from time to time, through maturation, education, or mishap. A *joint method trajectory* is a function:

$$\mathbf{c} : (G \times T) \to C,$$

that assigns a learning method $c \in C$ to each agent $i \in G$ at each time $t \in T$. A *possible world* is an arbitrary pair $w = (e_w, \mathbf{c}_w)$, such that $e_w \in E$ and each $\mathbf{c}_w$ is a joint method trajectory. Let $c_{i,w,t} = \mathbf{c}_w(i, t)$. Let W denote the set of all possible worlds.

A *preliminary computational learning model* (PCLM) for agents G is a quadruple $\mathfrak{M}_{t^*} = (E, \mathbf{s}, V, t^*)$ such that E is a non-empty set, $t^* \in T$ and:

$$\begin{aligned} \mathbf{s} &: (G \times W \times T) \to S; \\ V &: (\mathbf{L}_{\mathsf{atom}} \times T) \to \mathsf{Pow}(W). \end{aligned}$$

The function V is the usual *valuation function*, according to which $V(\mathsf{p}, t)$ is the proposition expressed by atomic sentence p at arbitrary time t. The distinguished time t^* is the

[6] Lower-case ϕ is also standardly employed in logic as a sentential in logical axiom schemata. Context readily disambiguates the two uses.

time of the epistemic context under discussion (Kamp 1971). Think of $s_{i,w,t} = \mathbf{s}(i,w,t)$ as the input that w presents to i at t in w. Call $\mathbf{s}$ the *input assignment function* and s_i the input assignment function for agent i. Define the *input stream* of i in w at t and the *input history* of i in w up to, but not including t as follows:

$$\begin{aligned} s_{i,w} &= (s_{i,w,0}, \ldots, s_{i,w,t}, \ldots); \\ s_{i,w}|t &= (s_{i,w,0}, \ldots, s_{i,w,t-1}). \end{aligned}$$

One major aim of this study is to provide a precise semantics for learnability, knowability, and the feasibility of knowing some things given that you know other things. It is assumed that changing the method of learner i does not cause changes to the external world or to the methods of the other agents. Therefore, the nearest world to w in which i uses method d at t is just the world $w[d/i,t]$ that results from substituting method d for agent i's method $c_{i,w,t}$ in w at t.[7]

Counterfactual shifts of method open the door to the medieval problem of information concerning future contingents, for since $\mathbf{s}(i,w,t)$ depends on w, which specifies i's method trajectory $c_{i,w}$, a crystal ball can send signals to i about the methods employed by i or other agents in the future, so counterfactual changes of method in the future could cause changes to past inputs. Learning semantics assumes that past inputs are preserved under future method choices. A *computational learning model* (CLM) is, accordingly, a PCLM that satisfies:

$$s_{i,w}|t = s_{i,w[d/i,t]}|t, \tag{1}$$

for all $i \in G$, $d \in C$, $w \in W$, and $t \in T$.

4. Information, Belief, and Determination

The input history $s_{i,w}|t$ of i in w has no truth value—it is a temporal sequence of sensory states—but it *makes available* to i in w at t the following, propositional *information*:[8]

$$\mathbf{I}(i,w,t) = \{w' \in W : s_{i,w}|t = s_{i,w'}|t\}.$$

In Kripke semantics for modal epistemic logic, available information is represented in terms of the *accessibility* relation "w' is possible in light of all the information available to i in w at t":

$$R_{i,t}(w,w') \Leftrightarrow w' \in I_{i,w,t}.$$

For fixed i and t, $(W, R_{i,t}, V)$ is a standard Kripke model. Since $R_{i,t}$ is an equivalence relation, the corresponding modal operator S5, as is often assumed (e.g., van Benthem 2010). Making propositional information available via physical signals is not the same thing as inserting that information directly into i's beliefs—it is still up to i's learning

[7] I.e.:

$$\begin{aligned} (\mathbf{c}[d/i,t])(i',t') &= \begin{cases} d & \text{if } i' = i \ \wedge \ t' = t; \\ \mathbf{c}(i',t') & \text{otherwise.} \end{cases} \\ w[d/i,t] &= (e_w, \mathbf{c}_w[d/i,t]). \end{aligned}$$

[8] Cf. (Lewis 1996) for a similar proposal.

function $L_{c_{i,w,t}}$ to interpret the signals, to recover the information they afford, and to incorporate it smoothly into i's belief system.

Possibilities of error that are incompatible with the information currently available will be deemed irrelevant to learning and knowledge. Furthermore, it does not seem that i needs to have been *informed* of her own learning method—the method merely has to *determine* success in light of available information. Accordingly, define the *determination assignment function*:[9]

$$\mathbf{D}(i, w, t) \quad = \quad \{w' \in I_{i,w,t} : c_{i,w,t^*} = c_{i,w',t^*}\}.$$

Then $D_{i,w,t} = \mathbf{D}(i, w, t)$ is the strongest proposition determined at t by the information and by the learning strategy possessed by i in w at t^*. The binary relation $\mathcal{D}_{i,t}(w, w')$ is again an equivalence relation that refines $\mathcal{I}_{i,t}(w, w')$.

Belief is handled very differently, as the concrete, hyper-intensional outcome of learning. None of the usual consistency, closure, or rationality assumptions is imposed, because they are all false. The actual belief state of i in w at t is produced by i's actual learning method at t:

$$\mathbf{B}_{\mathsf{act}}(i, w, t) \quad = \quad \{\phi \in \mathbf{L}_{\mathsf{@BIT}} \in \mathbf{L}_{\mathsf{@BIT}} : L_{c_{i,w,t}}(s_{i,w}|t, \phi) = 1\}.$$

In the long run, we are all dead and then we don't believe anything. An alternative account, in the spirit of (Nozick 1981), is that agent i would converge to true belief if *she* were to continue to use her current method forever:

$$\mathbf{B}_{\mathsf{ctr}}(i, w, t) \quad = \quad \{\phi \in \mathbf{L}_{\mathsf{@BIT}} : L_{c_{i,w,t^*}}(s_{i,w[c_{i,w,t^*}/i,t]}|t, \phi) = 1\}.$$

However, that would make it impossible for i to know inductively that all humans are mortal, since i would be immortal if she were literally to retain her current learning dispositions forever. Alternatively, one can focus on what i's current learning method *directs* i to believe in the future, just as one can speak of the outputs of an algorithm on inputs larger than any concrete machine running the algorithm will ever receive or of linguistic competence concerning sentences that will never be uttered due to resource limitations:

$$\mathbf{B}(i, w, t) \quad = \quad \{\phi \in \mathbf{L}_{\mathsf{@BIT}} : L_{c_{i,w,t^*}}(s_{i,w}|t, \phi) = 1 \ \wedge \ \phi \in \mathbf{L}_{\mathsf{BIT}}\}.$$

Refer to $B_{i,w,t} = \mathbf{B}(i, w, t)$ as the *virtual* belief state of i in w at t.

5. Learning Semantics

Let $\mathfrak{M}_{t^*} = (E, \mathbf{s}, V, t^*)$ be a CLM. Define the proposition $\|\phi\|^t_{\mathfrak{M}_{t^*}}$ expressed by ϕ in $\mathfrak{M}_{t^*}$ inductively as follows. In the base case:

$$\|\mathsf{p}\|^t_{\mathfrak{M}_{t^*}} \quad = \quad V(\mathsf{p}, t).$$

[9]This idea is also sketched in (Lewis 1996). The idea trivializes knowledge of one's own learning method. See the discussion in section 17.4 below for a potential, contextualist remedy. Also, it fails to rule out brain-in-a-vat worlds. See section 17.1 for a discussion of making determination safe or sensitive.

The connectives and $\perp$ have their standard, classical interpretations. For the temporal operators, define:

$$\begin{aligned}
\|\langle \mathsf{F}\rangle\, \phi\|^t_{\mathfrak{M}_{t^*}} &= \bigcup_{t' \geq t} \|\phi\|^{t'}_{\mathfrak{M}_{t^*}};\\
\|\langle \dot{\mathsf{F}}\rangle\, \phi\|^t_{\mathfrak{M}_{t^*}} &= \bigcup_{t' \geq t} \|\phi\|^{t'}_{\mathfrak{M}_{t'}};\\
\|\mathsf{N}\, \phi\|^t_{\mathfrak{M}_{t^*}} &= \|\phi\|^{t^*}_{\mathfrak{M}_{t^*}};\\
\|@_{t'}\, \phi\|^t_{\mathfrak{M}_{t^*}} &= \|\phi\|^{t'}_{\mathfrak{M}_{t^*}}.
\end{aligned}$$

Operator $\mathsf{F} = \langle \mathsf{F}\rangle$ is a future tense operator that includes the present time. Its dual is the "henceforth" operator G. Operator $\dot{\mathsf{F}} = \langle \dot{\mathsf{F}}\rangle$ is similar, except that it moves the epistemic context forward. Operator N resets the time t of evaluation to the time t^* of the epistemic context. Operator $@_{t'}$ resets the time of evaluation to the specified time t'.

Information and determination are defined propositionally, in the standard way, and both are S5 operators, for reasons already discussed.

$$\begin{aligned}
\|[\mathsf{I}]_i\, \phi\|^t_{\mathfrak{M}_{t^*}} &= \{w \in W:\ I_{i,w,t} \subseteq \|\phi\|^{t^*}_{\mathfrak{M}_{t^*}}\};\\
\|[\mathsf{D}]_i\, \phi\|^t_{\mathfrak{M}_{t^*}} &= \{w \in W: D_{i,w,t} \subseteq \|\phi\|^{t^*}_{\mathfrak{M}_{t^*}}\}.
\end{aligned}$$

Virtual belief, on the other hand, is entirely hyper-intensional, as it should be. Note that the time at which ϕ is believed to be true is always referred back to t^*, via the $@_{t^*}$ operator.

$$\|[\mathsf{B}]_i\, \phi\|^t_{\mathfrak{M}_{t^*}} = \{w \in W: @_{t^*}\, \phi \in B_{i,w,t}\}.$$

Methodological feasibility says that there is some method that i might have adopted that would achieve ϕ at t^* in w. It is used to express theses concerning learnability and knowability.

$$\|\langle \mathsf{M}\rangle_i\, \phi\|^t_{\mathfrak{M}_{t^*}} = \{w \in W: (\exists c \in C)\ w[c/i, t^*] \in \|\phi\|^t_{\mathfrak{M}_{t^*}}\}.$$

Methodological feasibility does not say that i can *guarantee* or *see to it* that ϕ is true. That stronger modality is expressed by $\langle \mathsf{M}\rangle_i [\mathsf{D}]_i$.

The remaining two operators are more subtle and work together as a team. To motivate conditional feasibility, consider the familiar logical thesis that the knowledge of i is closed under known consequence:

$$(\mathsf{K}_i\, \phi\ \wedge\ \mathsf{K}_i(\phi \to \psi)) \to \mathsf{K}_i\, \psi.$$

Granted, modus ponens is an easy inference to perform, but nothing like that thesis is even remotely true. Perhaps it is intended as a regulative ideal or as an obligation, but ideals are approachable and ought implies can, so the more proximate and concrete question is whether satisfaction of the thesis is *feasible*, in the sense that there is an *inference procedure* i could adopt that would *guarantee* that i knows that ψ *given* that she knows both ϕ and $\phi \to \psi$.

From the viewpoint of learning, effectively performing inferences amounts to an effective modification $h(c)$ of one's learning program c. Think of $\Delta \subseteq \mathbf{L}_{@\mathsf{BIT}}$ as a finite set of premises. One examines the verdicts of c for sentences in Δ (including, perhaps, past verdicts), and then one reverses the verdicts of c concerning some sentences (i.e., conclusions) outside of Δ. In that way, i can effectively modify her learning program c without having access either to c, itself, or to its raw, sub-cognitive, sensory inputs. Inference, therefore, makes sense as an evolutionary strategy—given some reptilian learning wet-ware that is hard to modify genetically without lethal effects, tack on some higher-level cognitive wet-ware that can intercept and modify its verdicts. That learning-theoretic conception of inference is made precise as follows. First, the *verdict* of learning method c concerning $\delta \in \mathbf{L}_{@\mathsf{BIT}}$ in response to σ is the ordered pair:

$$v_c(\sigma, \delta) \quad = \quad (L_d(\sigma, @_{t^*}\, \delta),\, L_d(\sigma, @_{t^*} \neg\delta)).$$

Define $c \equiv_\Delta d$ to hold if and only if learning programs c, d have identical verdicts for each $\delta \in \Delta$ and $\sigma \in S^*$. Define $c \equiv_{\Delta,\sigma} d$ to hold if and only if learning programs c, d have identical verdicts for each $\delta \in \Delta$ and for each initial segment τ of input history σ. Let h be a total recursive function that assumes values in C. Say that h *preserves* premises in Δ if and only if $h(c) \equiv_\Delta c$, for all $c \in C$. Say that h *depends only* on premises in Δ if and only if:

$$c \equiv_{\Delta,\sigma} d \;\Rightarrow\; v_{h(c)}(\sigma, \phi) = v_{h(d)}(\sigma, \phi).$$

for all $c, d \in C$, $\sigma \in S^*$, and $\phi \in \mathbf{L}_{@\mathsf{BIT}}$. Then h is an *inference procedure* with premises in Δ if and only if h is a total recursive function with range included in C that preserves premises in Δ and that depends only on premises in Δ.

Conditional feasibility expresses the existence of an inference procedure that *guarantees* the situation in the consequent, *given* the situation described in the antecedent. Accordingly, let $w \in \|\psi \,\langle\mathsf{MD}]\!\!\rightarrow_{i,\Delta}\, \phi\|^t_{\mathfrak{M}_{t^*}}$ if and only if there exists inference procedure h with premises in Δ such that, for all $u \in I_{i,w,t^*}$:

$$u \in \|\psi\|^t_{\mathfrak{M}_{t^*}} \quad \Rightarrow \quad u[h(c_{i,u,t})/i, t] \in \|\phi\|^t_{\mathfrak{M}_{t^*}}.$$

The notation $\psi \,\langle\mathsf{MD}]\!\!\rightarrow_{i,\Delta}\, \phi$ is mnemonic—existence of h is like $\langle\mathsf{M}\rangle_i$, the guarantee is like $[\mathsf{D}]_i$, and the assumption that the antecedent holds is like a conditional.

Inference—even deductive inference—can be subtly treacherous in learning semantics. Suppose that i contemplates changing her learning strategy c to d, which generates exactly the same verdict on δ that c does, in every possible input situation. Assumption (1) guarantees that d results in the same belief whether δ that c does *given the same inputs*, but the change from c to d could modify or even shut off the flow of *future* inputs to i because other agents detect the change in i (think of a poorly blinded social psychology experiment). Furthermore, the change from c to d could make δ false if the truth of δ depends on what some or all of the agents believe (e.g., i is a major player in the market). Either way, i's election to adopt inferential strategy d could be empirically or semantically *self-defeating*, in the sense that premise δ of the intended inference becomes untestable or false as a consequence of the inference being performed. Happily, good experimental design can prevent one's valid inferences from being self defeating, so it is useful to have vocabulary expressing that such preventive measures have successfully been carried out for some intended set of premises Δ. It is too strong to say that the inputs to i would be exactly the same whether i uses c or d, because i would presumably receive at least some information concerning her own beliefs. It suffices that neither the truth of the premises in Δ nor the verdicts of i concerning them is affected by the change. Define $w \in \|\mathsf{S}_i\, \Delta\|^t_{\mathfrak{M}_{t^*}}$

to hold if and only if for all $d \in C$ such that $c_{i,w,t^*} \equiv_\Delta d$ and for all $u \in D_{i,w,t^*}$, $t \geq t^*$, and $\delta \in \Delta$, if we set $u' = u[d/i, t^*]$ and $c = c_{i,u,t^*}(= c_{i,w,t^*})$, then:

$$\begin{aligned} u \in \|\delta\|^{t^*}_{\mathfrak{M}_{t^*}} &\Leftrightarrow u' \in \|\delta\|^{t^*}_{\mathfrak{M}_{t^*}}; \\ v_c(s_{i,u}|t, \delta) &= v_d(s_{i,u'}|t, \delta). \end{aligned}$$

That concludes the truth conditions for $\mathbf{L}_{\mathsf{BIT}}$. Let $\Gamma \subseteq \mathbf{L}_{\mathsf{BIT}}$. Define validity in a model and logical validity as follows:

$$\begin{aligned} \mathfrak{M}_{t^*} \models \phi &\Leftrightarrow W = \|\phi\|^{t^*}_{\mathfrak{M}_{t^*}}; \\ \models \phi &\Leftrightarrow \mathfrak{M}_{t^*} \models \phi, \text{ for each CLM } \mathfrak{M}_{t^*}. \end{aligned}$$

Note that validity in a model initializes time to the model's current epistemic context time t^*. Finally, logical entailment and equivalence are defined as follows:[10]

$$\begin{aligned} \phi \models \psi &\Leftrightarrow \models (\phi \rightarrow \psi); \\ \phi \equiv \psi &\Leftrightarrow \models (\phi \leftrightarrow \psi). \end{aligned}$$

6. Example: Outcomes of a Repeated Experiment

CLMs accommodate a boggling range of learning situations, but a collection of very elementary models suffices to illustrate many of the results that follow. Assume that each agent i passively observes the successive values of a repeated experiment whose outcomes are effectively coded as natural numbers. In the spirit of empiricism, identify possible external worlds with infinite outcome sequences $\varepsilon : \mathbb{N} \rightarrow \mathbb{N}$. Let E_0 denote the set of all such sequences. Define, for $k \in \mathbb{N}$:

$$\begin{aligned} \mathbf{s}_0(i, w, t) &= \varepsilon_w(t); \\ V_0(\mathsf{p}_k, t) &= \{\varepsilon \in E_0 : \varepsilon(t) = k\} \times \mathsf{C}^N; \\ \mathfrak{N}_{t^*} &= (E_0, \mathbf{s}_0, V_0, t^*). \end{aligned}$$

Temporal operators allow for compact expression of a range of increasingly complex statements:

$$\begin{aligned} \mathsf{p}_k &: \text{ the current outcome is } k; \\ \mathsf{G}\mathsf{p}_k &: \text{ the outcome will be } k; \\ \mathsf{F}\mathsf{p}_k &: \text{ the outcome is } k \text{ from now on}; \\ \mathsf{FG}\,\mathsf{p}_k &: \text{ the outcome will stabilize to value } k; \\ \mathsf{GF}\,\mathsf{p}_k &: \text{ the outcome is } k \text{ infinitely often.} \end{aligned}$$

A hypothesis ϕ is *objective* for i just in case i has the information available that i cannot alter the truth value of ϕ by changing her learning method. Objectivity simpliciter is objectivity for every agent.

$$\begin{aligned} \mathsf{O}_i\,\phi &:= [\,\mathsf{I}\,]_i(\phi \leftrightarrow [\mathsf{M}]_i\,\phi); \\ \mathsf{O}_G\,\phi &:= \bigwedge_{i \in G} \mathsf{O}_i\,\phi. \end{aligned}$$

[10]N.b. substitution of equivalents for equivalents under temporal operators does not preserve validity (Kamp 1971). For example, $\models \mathsf{G}(\phi \leftrightarrow \phi)$ and $\phi \equiv \mathsf{N}\,\phi$, but $\not\models \mathsf{G}(\phi \leftrightarrow \mathsf{N}\,\phi)$.

A special feature of model $\mathfrak{N}_{t^*}$ is that objectivity implies inferential stability:

$$\mathfrak{N}_{t^*} \models \mathsf{O}_i\,\phi \rightarrow \mathsf{S}_i\,\phi, \tag{2}$$

since inputs do not depend on methods at all, and neither does the truth of an objective statement.

7. Example: Agency, Games, and Experimentation

The agents in model $\mathfrak{N}_{t^*}$ are isolated natural scientists who passively receive inputs from a fixed experiment. But even a solipsistic scientist can choose how to interact with nature, and communication among scientists can produce cascades of interactive, doxastic effects. Although $\mathbf{L}_{\mathsf{BIT}}$ has no vocabulary describing acts other than belief, CLMs can represent arbitrarily complex social interactions involving such acts. The trick is to locate agents' diachronic strategies for non-doxastic actions within the "external world" $e \in E$. Then, all of the valid theses of learning semantics are valid for game-theoretic applications.

Here is one way to do it. Let $X \subseteq \mathbb{N}$ be a set of potential actions. Assuming that the actions are observable by all of the agents, let $S = X^N$. Then S^* contains all possible finite play histories. Let A denote the set of all $a \in \mathbb{N}$ such that ϕ_a is a unary total recursive function with range included in X. The *disposition to act* computed by a looks at the current input history and chooses how to act:

$$A_a(\sigma) \;=\; \phi_a(\langle\sigma\rangle).$$

Since belief depends on inputs, one special way for actions to depend on inputs is for them to depend on beliefs.

Dispositions to act can change through time just as dispositions to believe can. A *joint disposition trajectory* $\mathbf{a} : (G \times T) \rightarrow A$ assigns a profile of dispositions to the agents at each time. In purely social applications, the "external world" e can be identified with $\mathbf{a}$, so possible worlds are pairs $w = (\mathbf{a}, \mathbf{c})$. In experimental science, one agent can represent nature and the rest of the agents can be used to model socially distributed scientific inquiry. Each agent i receives as input the actions of every agent (including herself). Let $\sigma * s$ denote the concatenation of signal $s \in S$ to finite sequence $\sigma \in S^*$. The joint input assignment is then definable in stages as follows:

$$\begin{aligned}
\dot{s}_{i,(\mathbf{a},\mathbf{c})}|0 \;&=\; ();\\
\dot{s}_{i,(\mathbf{a},\mathbf{c})}|(t+1) \;&=\; \dot{s}_{i,(\mathbf{a},\mathbf{c})}|t * (A_{a_{i,t+1}}(\dot{s}_{i,(\mathbf{a},\mathbf{c})}|t) : i \leq N).
\end{aligned}$$

In the long run, all the players of an infinite game are dead, as are the dispositional properties of societies, economies, and terrestrial organisms. Hence, it is often more natural to think of the agents as *virtually* studying one another's and nature's *current* reactive dispositions, just as was done for belief:[11]

$$\begin{aligned}
s_{i,(\mathbf{a},\mathbf{c})}|t^* \;&=\; \dot{s}_{i,(\mathbf{a},\mathbf{c})}|t^*;\\
s_{i,(\mathbf{a},\mathbf{c})}|(t^*+t+1) \;&=\; s_{i,(\mathbf{a},\mathbf{c})}|t * (A_{a_{i,t^*}}(s_{i,(\mathbf{a},\mathbf{c})}|(t^*+t)) : i \leq N).
\end{aligned}$$

Either way, requirement (1) is satisfied.

In extensive form games, each agent receives some utility in each world at each time, as a result of what all the agents do. The utilities may also shift through time, if we

[11]The base case assumes that information gathered by means of earlier dispositions remains available.

interpret the agents as playing different games from time to time. Evolving utilities may be absorbed into the external world. The games described assume perfect information. Of course, $\mathbf{s}$ can easily be made to censor some actions.

8. Correctness and Error

Define "i is in *error* that ϕ" as follows:

$$\mathsf{E}_i\,\phi \;\;:=\;\; \mathsf{B}_i\,\phi \;\wedge\; \mathsf{N}\neg\phi.$$

Error *whether* ϕ is defined according to the general definition presented in section 2 above.

$$\tilde{\mathsf{E}}_i\,\phi \;\;:=\;\; \mathsf{E}_i\,\phi \;\vee\; \mathsf{E}_i\neg\phi.$$

It follows that i cannot be in error whether ϕ unless i believes that ϕ or believes that $\neg\phi$. That definition is straightforward if belief is deductively closed, but it is very weak for hyper-intensional belief—e.g., belief that ϕ does not count as an error whether $\neg\phi$. However, in order to interpret successful learning whether ϕ, all that is required is some unambiguous convention for i "getting ϕ wrong", and the proposed convention suffices in a minimal way. Stronger, but finite, demands on deductive acumen would not alter the results that follow, except to complicate their proofs. In a similar spirit, *correctness that* ϕ is absence of error whether ϕ together with belief that ϕ and correctness *whether* ϕ is defined as absence of error whether ϕ together with verdict for ϕ:

$$\begin{aligned}\mathsf{C}_i\,\phi \;\;&:=\;\; \neg\tilde{\mathsf{E}}_i\,\phi \;\wedge\; \mathsf{B}_i\,\phi;\\ \tilde{\mathsf{C}}_i\,\phi \;\;&:=\;\; \neg\tilde{\mathsf{E}}_i\,\phi \;\wedge\; \tilde{\mathsf{B}}_i\,\phi.\end{aligned}$$

Correctness whether ϕ could have been defined in the usual way as correctness that ϕ or correctness that $\neg\phi$, but that concept depends on whether i believes that $\neg\neg\phi$. The proposed definition depends only on i's belief whether ϕ.[12]

9. Inductive Learning

In computational learning theory, *inductive learning* whether ϕ is understood as guaranteed convergence of i's current learning method to correct belief whether ϕ. That is elegantly formalizable in $\mathbf{L}_{\mathsf{BIT}}$ as follows:

$$\tilde{\mathsf{L}}_i\,\phi \;\;:=\;\; [\mathsf{D}]_i\mathsf{F}\mathsf{G}\tilde{\mathsf{C}}_i\,\phi.$$

The truth conditions for $\tilde{\mathsf{L}}_i\,\phi$ can be expressed entirely in terms of the proposition $\|\phi\|^{t^*}_{\mathfrak{M}_{t^*}}$ and the verdicts of i's learning method: $w \in \|\mathsf{L}_i\,\phi\|^{t}_{\mathfrak{M}_{t^*}}$ if and only if for all $u \in D_{i,w,t}$,

$$u \in \|\phi\|^{t^*}_{\mathfrak{M}_{t^*}} \;\;\Rightarrow\;\; \big(\lim_{t\to\infty} L_{c_{i,w,t^*}}(s_{i,u}|t, @_{t^*}\,\phi) \;=1 \;\wedge\; \lim_{t\to\infty} L_{c_{i,w,t^*}}(s_{i,u}|t, @_{t^*}\,\neg\phi) = 0\big); \tag{3}$$

$$u \notin \|\phi\|^{t^*}_{\mathfrak{M}_{t^*}} \;\;\Rightarrow\;\; \big(\lim_{t\to\infty} L_{c_{i,w,t^*}}(s_{i,u}|t, @_{t^*}\,\neg\phi) = 1 \;\wedge\; \lim_{t\to\infty} L_{c_{i,w,t^*}}(s_{i,u}|t, @_{t^*}\,\phi) \;=0\big). \tag{4}$$

[12]Thanks to Ted Shear for this point.

That is essentially equivalent to saying, in computational learning theory, that i's *current* method c_{i,w,t^*} *decides* ϕ in the limit (Kelly 1996), except that learning semantics allows that the data depend on the learning method.

10. Inductive Learnability

Just as the theory of computability concerns what can be computed, rather than how we actually compute, computational learning theory focuses on learnability—the feasibility of learning—rather than on the actual psychology of learning. Learning semantics affords at least four grades of feasibility:

(5) $$\langle \mathsf{M} \rangle_i [\mathsf{D}]_i \phi \quad \models \quad [\mathsf{D}]_i \langle \mathsf{M} \rangle_i \phi \quad \models \quad \langle \mathsf{M} \rangle_i \phi \quad \models \quad \langle \mathsf{M} \rangle_i \langle \mathsf{D} \rangle_i \phi.$$

In the case of learnability, those concepts collapse to $\langle \mathsf{M} \rangle_i \tilde{\mathsf{L}}_i \phi$—the last entails the first, since $\tilde{\mathsf{L}}_i$ begins with $[\mathsf{D}]_i$, which is an S5 operator:

(6) $$\langle \mathsf{M} \rangle_i [\mathsf{D}]_i \tilde{\mathsf{L}}_i \phi \quad \equiv \quad [\mathsf{D}]_i \langle \mathsf{M} \rangle_i \tilde{\mathsf{L}}_i \phi \quad \equiv \quad \langle \mathsf{M} \rangle_i \tilde{\mathsf{L}}_i \phi \quad \equiv \quad \langle \mathsf{M} \rangle_i \langle \mathsf{D} \rangle_i \tilde{\mathsf{L}}_i \phi.$$

Concretely, $w \in \| \langle \mathsf{M} \rangle_i \tilde{\mathsf{L}}_i \phi \|^t_{\mathfrak{M}_{t^*}}$ if and only if there exists $d \in C$ such that (3) and (4) hold with d substituted for c_{i,w,t^*} in u, for all $u \in I_{i,w[d/i,t^*],t}$. If ϕ satisfies $\mathsf{O}_i \phi$ in $\mathfrak{N}_{t^*}$, one can also substitute $I_{i,w,t}$ for $I_{i,w[d/i,t^*],t}$, in which case the truth conditions for learnability are essentially the same as the conditions for decidability in the limit (Kelly 1996).[13]

Universal truths and existential truths about the future are inductively learnable in the empirical model $\mathfrak{N}_{t^*}$—just believe the universal hypothesis until it is refuted and believe its negation thereafter, and follow the dual strategy in the existential case:

(7) $$\mathfrak{N}_{t^*} \quad \models \quad \langle \mathsf{M} \rangle_i \tilde{\mathsf{L}}_i \mathsf{G}\, \mathsf{p}_k;$$

(8) $$\mathfrak{N}_{t^*} \quad \models \quad \langle \mathsf{M} \rangle_i \tilde{\mathsf{L}}_i \mathsf{F}\, \mathsf{p}_k.$$

But not every empirical hypothesis is inductively learnable. Kant (1782/1787) observed that hypotheses like the finite or infinite divisibity of matter or the existence of a first moment in time "outpace all possible experience". In terms of learnablity, he was right. Suppose that the laboratory returns a 1 whenever an allegedly fundamental particle is split and returns a 0 when an attempted split fails. Then finite divisibility of matter can be expressed as $\mathsf{FG}\, \mathsf{p}_0$ and infinite divisibility of matter can be expressed as $\mathsf{GF}\, \mathsf{p}_1$. Both hypotheses are evidently only remotely connected with current experience. In fact, neither is inductively learnable in $\mathfrak{N}_{t^*}$:

(9) $$\mathfrak{N}_{t^*} \quad \models \quad \neg \tilde{\mathsf{L}}_i \mathsf{FG}\, \mathsf{p}_k;$$

(10) $$\mathfrak{N}_{t^*} \quad \models \quad \neg \tilde{\mathsf{L}}_i \mathsf{GF}\, \mathsf{p}_k.$$

It suffices to show, via a standard, learning theoretic diagonal argument, that no c satisfies convergence conditions (3) and (4).[14]

Learning semantics is a flexible framework for inductive learning and learnability that allows one, for the first time, to rigorously iterate the learning operator, in order to analyze precisely such statements as that it is learnable whether someone else is learning whether ϕ. But in order to provide the sharpest possible contrast between learning semantics and

[13] The differences concern mere conventions for coding the acceptance, rejection, or suspension of belief of i with respect to ϕ.

[14] Proofs of selected theses are presented in the appendix.

traditional possible worlds semantics for epistemic logic, the focus of this study is on the semantics of inductive knowledge, to which we now turn.

11. Inductive Knowledge

Agent i *has learned* whether [that] ϕ if and only if i is learning whether ϕ and, henceforth, i correctly (virtually) believes whether [that] ϕ:

$$\begin{aligned} \tilde{\mathsf{Led}}_i\,\phi &:= \mathsf{G}\tilde{\mathsf{C}}_i\,\phi \ \wedge\ \tilde{\mathsf{L}}_i\,\phi; \\ \mathsf{Led}_i\,\phi &:= \mathsf{G}\mathsf{C}_i\,\phi \ \wedge\ \tilde{\mathsf{L}}_i\,\phi \ \equiv\ \tilde{\mathsf{Led}}_i\,\phi \ \wedge\ \phi. \end{aligned}$$

Having learned inductively whether ϕ may sound odd, since the culmination of inductive inquiry depends on what i's current learning method would do in the future. But such locutions are actually quite common: e.g., "I have quit smoking for good".

It is natural to suppose that inductive knowledge is having learned, but there is a powerful argument to the contrary: learnability is not preserved under logical consequence; for recall (7), (9), and (10) and note that $\mathsf{G}\,\phi$ entails both $\mathsf{GF}\,\phi$ and $\mathsf{FG}\,\phi$. Since having learned entails learnability, it follows that knowability is not closed under logical consequence. And the examples sound bad: we would know that the laws of quantum mechanics apply invariably, but it would be unknowable that they apply infinitely often or all but finitely often. It sounds better to say that we know that a prethe latter two statements *because* we know the first.

Pursuing that idea, suppose that i's only reason for believing that $\mathsf{GF}\,\phi$ is that she believes $\mathsf{G}\,\phi$ and suppose that her reason for believing $\mathsf{G}\,\phi$ is that it has stood up to severe testing so far (a single counterexample would refute $\mathsf{G}\,\phi$). It is a traditional theme in the philosophy of science that general theories are not testable until they are *articulated* with auxiliary assumptions (Duhem 1914). Semantically speaking, "articulation" amounts to the substitution of a logically stronger, testable hypothesis for the untestable hypothesis, itself. Thus, one may think of $\mathsf{G}\,\phi$ as a testable articulation of $\mathsf{GF}\,\phi$, since it posits a particularly simple *way* in which $\mathsf{GF}\,\phi$ *might* be true. Then i stabilizes to true belief that $\mathsf{GF}\,\phi$ as soon as i stabilizes to true belief that $\mathsf{G}\,\phi$, so the actual convergence requirement is met also for $\mathsf{GF}\,\phi$. But what if $\mathsf{G}\,\phi$ were to be refuted, say at time t? Maybe i has plausible ideas about how to re-articulate $\mathsf{GF}\,\phi$ (e.g., as $@_{t+1}\mathsf{G}\,\phi$). In order to learn by such a strategy, i would require a full contingency plan for re-articulating $\mathsf{GF}\,\phi$ that somehow hits upon a true articulation eventually in *every* possible world in which $\mathsf{GF}\,\phi$ is true. But it has already been shown that no such contingency plan exists for $\mathsf{GF}\,\phi$, since $\mathsf{GF}\,\phi$ is not learnable.

Another venerable theme in the philosophy of science is that there is "no logic of discovery" (Hempel 1945, Popper 1935), which means, roughly, that science need not have an explicit contingency plan for what to propose when old hypotheses are refuted, so far as scientific knowledge is concerned. The standard arguments for that conclusion are analogical and historical.[15] The argument from analogy is that a theorem is still a theorem no matter how one came to conjecture it, so scientific knowledge likewise does not depend on how one came to think up the hypothesis. The historical argument is that major scientific findings have been hit upon by luck. For a celebrated example, the chemist Kekulé

[15] A notable exception is (Putnam 1963).

claimed to discover the carbon ring structure of benzine by dreaming of a snake biting its tail (Hempel 1945, Benfey 1958). It does not seem to count against Kekulé's subsequent knowledge of that hypothesis that he possessed no systematic contingency plan for dreaming up alternative molecular structures, had the ring hypothesis failed. Scientists refer to luck that does not undermine scientific knowledge as *serendipity.* Kekule's dream was serendipitous in that sense, as is all luck in hitting upon a true hypothesis. Since untestable hypotheses like $\mathsf{GF}\,\phi$ cannot be learned, they can be known only with serendipity. So allowance for serendipity, the practice of testing testable articulations of untestable hypotheses, and the slogan that there is "no logic of discovery" are all grounded in the closure of inductive knowability under logical consequence, a fundamental, epistemological consideration.

Suppose that i is commanded by her thesis advisor to investigate $\mathsf{GF}\,\phi$ by severely testing $\mathsf{G}\,\phi$. We know that i lacks a full logic of discovery for $\mathsf{GF}\,\phi$, since $\mathsf{GF}\,\phi$ is not learnable. Suppose, plausibly, that she has far less—if $\mathsf{G}\,\phi$ is ever refuted, she has no idea what is going on, suspends belief forever whether $\mathsf{GF}\,\phi$, and switches to a more lucrative career in finance. *If* her advisor was right (serendipity), then she has already converged to true belief that $\mathsf{GF}\,\phi$ and, since her belief that $\mathsf{GF}\,\phi$ is based *solely* on her belief that $\mathsf{G}\,\phi$, she is also guaranteed to eliminate error with respect to $\mathsf{GF}\,\phi$ eventually. Her (actual) convergence to true belief that the untestable hypothesis is true is serendipitous, but her eventual avoidance of error is not lucky at all—it is guaranteed by her commitment to suspend belief forever if $\mathsf{GF}\,\phi$ is refuted.

In light of the preceding considerations, it is proposed that inductive knowledge that ϕ is actual convergence to true belief that ϕ along with guaranteed, eventual avoidance of error whether ϕ:[16]

$$\begin{aligned}\tilde{\mathsf{K}}_i\,\phi \;&:=\; \mathsf{G}\tilde{\mathsf{C}}_i\,\phi \;\wedge\; [\mathsf{D}]_i\mathsf{FG}\neg\tilde{\mathsf{E}}_i\,\phi;\\ \mathsf{K}_i\,\phi \;&:=\; \mathsf{G}\mathsf{C}_i\,\phi \;\wedge\; [\mathsf{D}]_i\mathsf{FG}\neg\tilde{\mathsf{E}}_i\,\phi \;\equiv\; \tilde{\mathsf{K}}_i\,\phi \;\wedge\; \phi.\end{aligned}$$

Thus, having learned whether ϕ is sufficient, but not necessary, for knowing whether ϕ:

(11) $$\models\; \tilde{\mathsf{Led}}_i\,\phi \rightarrow \tilde{\mathsf{K}}_i\,\phi;$$

(12) $$\models\; \mathsf{Led}_i\,\phi \rightarrow \mathsf{K}_i\,\phi.$$

[16]Hendricks (2001) presents several concepts of empirical knowledge, the closest of which to the following proposal is "realistic reliable true belief" or *RRT knowledge.* Hendricks' informal gloss of RRT knowledge (p. 181) amounts to the following proposal in the present notation: $\mathsf{Krrt}_i\,\phi := \mathsf{G}\,\phi \;\wedge\; \tilde{\mathsf{L}}_i\,\phi$ (the operator $[\mathsf{D}]_i$ is dropped from the $\tilde{\mathsf{L}}_i\,\phi$ condition in the accompanying formal statement—presumably unintentionally). RRT knowledge is very different from inductive knowledge as defined here. First of all, RRT knowledge requires that $\mathsf{G}\,\phi$, which would make it impossible for i to know, for example, that she believes that ϕ, if that belief state is transient. Learning semantics sidesteps that difficulty by evaluating the proposition believed at the "now" of utterance. Second, RRT knowledge does not require $\mathsf{GB}_i\,\phi$, so RRT knowledge does not even imply belief that ϕ, much less stable belief that ϕ—it may be years until the learning process succeeds. Finally, RRT knowledge does imply learning whether ϕ, which implies that RRT knowability cannot be closed under deductive consequence, as has just been explained. Hendricks' claim that RRT knowledge validates the axioms of modal system S4 (proposition 12.3, p. 208) is therefore false. The discrepancy is explained by the fact that, just prior to the proof of proposition 12.3, Hendricks inadvertently modifies the concept of RRT knowledge a second time (p. 194) to $\mathsf{G}\,\phi$ conjoined with the existence of a future time t' such that it is determined now that i believes that ϕ forever after t'—whether or not ϕ is true.

In fact, learning is *equivalent* to guaranteed, eventual arrival at knowledge—a nice example of a plausible validity expressible in $\mathbf{L}_{\mathsf{BIT}}$ but not in the traditional, pure K_i fragment.[17]

$$\tilde{\mathsf{L}}_i\, \phi \;\;\equiv\;\; [\mathsf{D}]_i \mathsf{F} \tilde{\mathsf{K}}_i\, \phi. \tag{13}$$

In terms of concrete learning methods, the first conjunct of $\tilde{\mathsf{K}}\, \phi$ is true in w at t if and only if:

$$w \in \|\phi\|^{t^*}_{\mathfrak{M}_{t^*}} \;\Rightarrow\; ((\forall t \geq t^*)\; L_{c_{i,w,t^*}}(s_{i,w}|t, @_{t^*}\, \phi) \;= 1 \;\wedge \tag{14}$$

$$(\forall t \geq t^*)\; L_{c_{i,w,t^*}}(s_{i,w}|t, @_{t^*}\, \neg\phi) = 0);$$

$$w \notin \|\phi\|^{t^*}_{\mathfrak{M}_{t^*}} \;\Rightarrow\; ((\forall t \geq t^*)\; L_{c_{i,w,t^*}}(s_{i,w}|t, @_{t^*}\, \phi) \;= 0 \;\wedge \tag{15}$$

$$(\forall t \geq t^*)\; L_{c_{i,w,t^*}}(s_{i,w}|t, @_{t^*} \neg\phi) \;= 1);$$

and the second conjunct is true in w at t if and only if for all $u \in I_{i,w,t}$:

$$u \in \|\phi\|^{t^*}_{\mathfrak{M}_{t^*}} \;\Rightarrow\; \lim_{t\to\infty} L_{c_{i,w,t^*}}(s_{i,u}|t, @_{t^*}\, \neg\phi) = 0; \tag{16}$$

$$u \notin \|\phi\|^{t^*}_{\mathfrak{M}_{t^*}} \;\Rightarrow\; \lim_{t\to\infty} L_{c_{i,w,t^*}}(s_{i,u}|t, @_{t^*}\, \phi) \;= 0. \tag{17}$$

Note that (16) and (17) weaken the corresponding conditions (3) and (4) for having learned.

12. Inductive Knowability

Learning semantics again affords the following notions of inductive knowability, in descending strength:

$$\langle \mathsf{M} \rangle_i [\mathsf{D}]_i \mathsf{K}_i\, \phi \quad\models\quad [\mathsf{D}]_i \langle \mathsf{M} \rangle_i \mathsf{K}_i\, \phi \quad\models\quad \langle \mathsf{M} \rangle_i \mathsf{K}_i\, \phi \quad\models\quad \langle \mathsf{M} \rangle_i \langle \mathsf{D} \rangle_i \mathsf{K}_i\, \phi. \tag{18}$$

The four conditions of knowability are all logically distinct. The first version requires correct belief immediately. implies a sweeping inductive skepticism, in the sense that ϕ is inductively knowable by i only if the information is available to i that i can make ϕ true:

$$\models \langle \mathsf{M} \rangle_i [\mathsf{D}]_i \mathsf{K}_i\, \phi \rightarrow [\,\mathsf{I}\,]_i \langle \mathsf{M} \rangle_i\, \phi. \tag{19}$$

The trouble is that the But the remaining versions do allow for inductive knowability. For example:

$$\models [\mathsf{D}]_i \langle \mathsf{M} \rangle_i \mathsf{K}_i \mathsf{G} \mathsf{p}_k. \tag{20}$$

The i

That leaves the weakest option, which requires only that it be feasible for i to make it *possible* that she knows now—an idea consonant with serendipity:

$$\langle \mathsf{MD} \rangle_i\, \phi \;:=\; \langle \mathsf{M} \rangle_i \langle \mathsf{D} \rangle_i \tilde{\mathsf{K}}_i\, \phi \tag{21}$$

$$\equiv\; \langle \mathsf{M} \rangle_i \langle \mathsf{D} \rangle_i (\mathsf{G} \tilde{\mathsf{C}}_i\, \phi \;\wedge\; [\mathsf{D}]_i \mathsf{F} \mathsf{G} \neg \tilde{\mathsf{E}}_i\, \phi) \tag{22}$$

$$\equiv\; \langle \mathsf{M} \rangle_i (\langle \mathsf{D} \rangle_i \mathsf{G} \tilde{\mathsf{C}}_i\, \phi \;\wedge\; [\mathsf{D}]_i \mathsf{F} \mathsf{G} \neg \tilde{\mathsf{E}}_i\, \phi); \tag{23}$$

[17]Thesis (13) is invalid with $\dot{\mathsf{F}}$ in place of F. It is crucial that the doxastic future under consideration is virtual rather than actual.

where the last equivalence is again due to $[\mathsf{D}]_i$ being S5. Condition (23) expands to the existence of $d \in C$ such that for some $u \in I_{i,w,t}$:

$$u[d/i,t] \in \|\phi\|^t_{\mathfrak{M}_{t^*}} \;\Rightarrow\; ((\forall t \geq t^*)\ L_d(s_{u,i}|t, @_{t^*}\,\phi) = 1 \;\wedge\; (\forall t \geq t^*)\ L_d(s_{u,i}|t, @_{t^*}\,\neg\phi) = 0); \tag{24}$$

$$u[d/i,t] \notin \|\phi\|^t_{\mathfrak{M}_{t^*}} \;\Rightarrow\; ((\forall t \geq t^*)\ L_d(s_{u,i}|t, @_{t^*}\,\phi) = 0 \;\wedge\; (\forall t \geq t^*)\ L_d(s_{u,i}|t, @_{t^*}\,\neg\phi) = 1); \tag{25}$$

and for all $u \in I_{i,w,t}$:

$$u[d/i,t] \in \|\phi\|^t_{\mathfrak{M}_{t^*}} \;\Rightarrow\; \lim_{t\to\infty} L_d(s_{i,u}|t, @_{t^*}\,\neg\phi) = 0; \tag{26}$$

$$u[d/i,t] \notin \|\phi\|^t_{\mathfrak{M}_{t^*}} \;\Rightarrow\; \lim_{t\to\infty} L_d(s_{i,u}|t, @_{t^*}\,\phi) = 0. \tag{27}$$

Conditions (24) and (25) are trivially satisfiable by dogmatically believing that ϕ and conditions (26) and (27) are trivially satisfiable by skeptically suspending belief whether ϕ. But the conditions are not jointly trivial—the possibility of having converged to the truth risks the possibility of error infinitely often, unless one has an appropriate plan in place for when to suspend judgment, as Popper (1935) insisted. For example, weak knowability can fail when even the total input stream does not determine the truth of ϕ in any world. In that case, say that ϕ is *globally underdetermined*—venerable candidates include "the Absolute is lazy" and Poincare's (1904) perfect trade-off between shrinking forces and geometry. The logical positivists attempted to rule out globally underdetermined hypotheses by deeming them meaningless, on empiricist grounds, but that leaves open the question whether freedom from global underdetermination guarantees knowability. Learning semantics validates something close to that in the empiricist model $\mathfrak{N}_{t^*}$, as long as the input stream is computable. Recall the strategy, discussed above, of guessing a testable articulation ψ of ϕ, believing ϕ until ψ is refuted, and suspending judgment thereafter. It witnesses the following, liberal knowability condition for objective hypotheses in $\mathfrak{N}_{t^*}$:

Proposition 1. *Suppose that* $w \in \|\mathsf{O}_G\,\phi\|^{t^*}_{\mathfrak{N}_{t^*}}$ *and there exists* $u \in I_{i,w,t^*} \cap \|\phi\|^{t^*}_{\mathfrak{N}_{t^*}}$ *with computable input stream* $s_{i,u}$. *Then* $w \in \|\langle\mathsf{MD}\rangle_i\ \mathsf{K}_i\,\phi\,\|^{t^*}_{\mathfrak{N}_{t^*}}$.

As a corollary, we have the following knowability result, in contrast to the non-learnability results (9) and (10) above:[18]

$$\mathfrak{N}_{t^*} \;\models\; \langle\mathsf{MD}\rangle_i(\mathsf{K}_i\mathsf{G}\,\mathsf{p}_k \;\wedge\; \mathsf{K}_i\mathsf{F}\,\mathsf{p}_k \;\wedge\; \mathsf{K}_i\mathsf{FG}\,\mathsf{p}_k \;\wedge\; \mathsf{K}_i\mathsf{GF}\,\mathsf{p}_k). \tag{28}$$

The restriction to $\mathfrak{N}_{t^*}$ and to objective ϕ rules out global underdetermination. The assumption that $s_{i,w}$ is computable is also crucial. For example, take the setting to be $\mathfrak{N}_{t^*}$ restricted to worlds that present binary data. Add a new atomic sentence q with the valuation $V(\mathsf{q}) = \{w \in W : s_{i,u} = g\}$, where g is a fixed, total, non-computable, binary-valued function. Call the resulting model $\mathfrak{B}_{t^*}$. Then we have:[19]

$$w \notin \|\langle\mathsf{MD}\rangle_i\mathsf{K}_i\,\mathsf{q}\|^{t^*}_{\mathfrak{B}_{t^*}}. \tag{29}$$

[18] Just let u satisfy $s_{i,u,t} = s_{i,w,t}$ for $t < t^*$ and $s_{i,u,t} = k$ for $t \geq t^*$.

[19] The restriction to binary sequences in (29) matters. If the range of inputs at each stage might be infinite, then one can add an atomic sentence to $\mathfrak{N}_{t^*}$ that is knowable but true only in worlds that are empirically *infinitely* uncomputable (cf. Kelly 1996, 7.19).

This short introduction to the logic of inductive knowability illustrates that the proposed semantics focuses attention precisely where it should—on concrete, methodological considerations like computability and global underdetermination. Furthermore, it is of interest that allowance for serendipitous knowledge allows not only for deductive closure of knowability, but also for a considerable broadening of the scope of inductive knowability beyond that of learnability.[20]

13. Fitch's Paradox

It has just been shown that, in learning semantics, the question of inductive knowability raises concrete, familiar, methodological issues. Since traditional epistemic logic makes no contact with learning, either in its syntax or in its models, it focuses on the more exotic problem of unknowability due to epistemic self-reference. Although self-referential paradoxes are remote from the concrete business of science, questions of genuine epistemological interest, such as whether it is possible for science to know inductively that it does not know inductively, open the logical floodgates to self-referential curiosities. Alas, one cannot simply ignore them. At the very least, one must construct a firewall against them that does not trivialize the principles of interest.

Consider, for example, the *Moore sentence* for ϕ, defined as follows:

$$\mathsf{Mo}_i\,\phi \;\; := \;\; \phi \;\wedge\; \neg \mathsf{K}_i\,\phi.$$

The Moore sentence is not knowable in standard epistemic logic, for suppose that i knows that $\mathsf{Mo}_i\,\phi$. Then, since knowledge is true, $\mathsf{Mo}_i\,\phi$ is also be true, so $\neg\mathsf{K}_i\,\phi$ is true. But since $\mathsf{Mo}_i\,\phi$ is known, so is conjunct ϕ of $\mathsf{Mo}_i\,\phi$, so $\mathsf{K}_i\,\phi$ is true. Contradiction. The proof requires only (i) that the conjuncts of a known conjunction are known and (ii) that knowledge is true, both of which are valid in standard, possible world semantics.

That is hardly surprising in itself, but it leads directly[21] to *Fitch's paradox*, the statement that any agent for whom every truth ϕ is knowable is *already omniscient*.[22]

$$(\forall\phi)\;(\phi \rightarrow \Diamond_i \mathsf{K}_i\,\phi) \rightarrow (\forall\phi)\;(\phi \rightarrow \mathsf{K}_i\,\phi). \tag{30}$$

For suppose that the consequent of (30) is false. Then $(\exists\phi)\;\mathsf{Mo}_i\,\phi$ is true. But $\mathsf{Mo}_i\,\phi$ is not knowable. So $\mathsf{Mo}_i\,\phi$ is a counterexample to the antecedent of (30).

Fitch's paradox is not really paradoxical after the "gotcha" moment when one realizes that denying the consequent yields a true Moore sentence. If "every truth" is restricted to "every scientifically interesting, objective truth", the paradox evaporates. Nonetheless, there is a specialist literature devoted to refuting Fitch's paradox, some authors going so far as to blame proof by contraposition (Williamson 1993). Therefore, it may be of interest to revisit the question whether the Moore sentence is knowable in learning semantics. The standard argument that $\mathsf{Mo}_i\,\phi$ is not knowable assumes that i knows the conjuncts of any conjunction i knows. That step evidently fails in learning semantics, because even belief is not closed under deductive consequence. But inferring ϕ, ψ from $\phi \wedge \psi$ is the easiest of

[20] Cf. sub-section 14.1 below for a formal discussion of deductive closure of inductive knowledge.

[21] The ingenious step was taken by Alonzo Church (2009) in an anonymous referee report on Fitch's manuscript.

[22] It suffices that $\Diamond_i$ be the dual of an alethic necessity operator satisfying the rule of necessitation.

inferences—one need only erase the $\wedge$. It would, therefore, be more sporting to show that $\mathsf{Mo}_i\,\phi$ is knowable by an agent whose beliefs are *conjunctively cogent* in the sense that:

$$\mathsf{Coco}_i(\phi, \psi) \;\; := \;\; [\,\mathsf{I}\,]_i(\mathsf{B}_i(\phi \;\wedge\; \psi) \leftrightarrow (\mathsf{B}_i\,\phi \;\wedge\; \mathsf{B}_i\,\psi)).$$

Learning semantics yields a startling, positive verdict:[23]

$$(31) \qquad \mathfrak{N}_{t^*} \;\models\; (\mathsf{O}_i\,\phi \;\wedge\; \neg[\mathsf{D}]_i\,\phi \;\wedge\; \langle\mathsf{MD}\rangle_i \mathsf{K}_i\,\phi) \;\rightarrow\; \rightarrow \langle\mathsf{MD}\rangle_i(\mathsf{K}_i\mathsf{Mo}_i\,\phi \;\wedge\; \mathsf{Coco}_i(\phi, \neg\mathsf{K}_i\,\phi)).$$

Of course, some sort of aphasia is required to know one's own Moore sentence, but the aphasia now plausibly concerns learning, rather than a trivial, deductive inference.[24] Suppose that i is irrecoverably dogmatic that ϕ. When an acquaintance accuses i of not knowing that ϕ, even though ϕ is true (the evidence for ϕ is abundant), i takes a detached interest in the accusation. Since i's admitted dogmatism precludes her from knowing that ϕ, the knowability of $\mathsf{Mo}_i\,\phi$ reduces, for i, to that of ϕ, so i can know $\mathsf{Mo}_i\,\phi$ by basing her belief whether $\mathsf{Mo}_i\,\phi$ on the evidence concerning ϕ. Since ϕ is knowable, $\mathsf{Mo}_i\,\phi$ is knowable by i—because i is dogmatic with respect to ϕ. Furthermore, due to i's dogmatism, i is conjunctively cogent with respect to the conjuncts of $\mathsf{Mo}_i\,\phi$. We may not envy i's strange knowledge, but the story is plausible enough and standard possible world semantics for K_i and for $\langle\mathsf{MD}\rangle_i$ cannot accommodate it, much less help one to discover it. A general moral for modal epistemic semantics is that abstraction from the details of inquiry provides no guarantee against philosophical error.

The preceding discussion notwithstanding, learning models still permit one to construct self-referential monstrosities by "brute force", using the valuation function: e.g., an atomic sentence can be interpreted to say "i does not believe that she knows me". Such models trivially invalidate the thesis that it is feasible to know that one knows what one knows. The real purpose of the doxastic stability operator S_i is to protect otherwise plausible theses of epistemic logic from that self-referential onslaught. Under the hypothesis that $\mathsf{S}_i\,\phi$ obtains, knowledge, learning and having learned are preserved under counterfactual changes of method that do not modify the agent's current learning disposition with respect to ϕ.[25]

Proposition 2. *Suppose that* $u \in \|\mathsf{S}_i\,\Delta\|^{t^*}_{\mathfrak{M}_{t^*}}$ *and* $d \equiv_\Delta c_{i,u,t^*}$ *and* $\phi \in \Delta$. *Then:*

$$(32) \qquad u \in \|\mathsf{K}_i\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}} \;\;\Rightarrow\;\; u[d/i, t^*] \in \|\mathsf{K}_i\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}};.$$

and similarly for $\tilde{\mathsf{K}}_i$, $\tilde{\mathsf{L}}_i$, $\tilde{\mathsf{Led}}_i$ *and* Led_i.

[23]Note that there is no temporal equivocation here between the time at which $\mathsf{Mo}_i\,\phi$ is known and the time at which ϕ is not known, as there is in solutions proposed in temporal dynamic epistemic logic (e.g., Yap and Hoshi 2009).

[24]Alternative learning strategies within the same agent are a familiar theme in the epistemology literature—e.g., (Nozick 1981).

[25]In the author's opinion, finding a semantics for S_i such that $\mathsf{S}_i\,\phi$ is both plausible and yet strong enough to yield the following invariance property proved to be the crux of the entire subject.

14. Epistemic Logic Redux

The idea in traditional epistemic logic is to mine intuitions for principles stated entirely in terms of K_i and then to solve backwards for conditions on the accessibility relation that validate them. Modal semantics then serves as a silent bookkeeper that faithfully manages the iteration of K_i, subject to those assumptions.[26] Here is a standard menu of potential principles to impose:

$$\begin{array}{lll}
\mathrm{N} & : & \mathsf{K}_i\,\phi, \text{ if } \models \phi; \\
\mathrm{K} & : & \mathsf{K}_i(\phi \to \psi) \to (\mathsf{K}_i\,\phi \to \mathsf{K}_i\,\psi); \\
\mathrm{T} & : & \mathsf{K}_i\,\phi \to \phi; \\
\mathrm{B} & : & \phi \to \mathsf{K}_i\neg\mathsf{K}_i\neg\phi; \\
4 & : & \mathsf{K}_i\,\phi \to \mathsf{K}_i\mathsf{K}_i\,\phi; \\
0.2 & : & \neg\mathsf{K}_i\neg\mathsf{K}_i\,\phi \to \mathsf{K}_i\neg\mathsf{K}_i\neg\phi; \\
0.3 & : & \mathsf{K}_i(\mathsf{K}_i\,\phi \to \mathsf{K}_i\,\psi) \vee \mathsf{K}_i(\mathsf{K}_i\,\psi \to \mathsf{K}_i\,\phi); \\
0.4 & : & \phi \to (\neg\mathsf{K}_i\neg\mathsf{K}_i\phi \to \neg\mathsf{K}_i\,\phi); \\
5 & : & \neg\mathsf{K}_i\,\phi \to \mathsf{K}_i\neg\mathsf{K}_i\,\phi.
\end{array}$$

For a happy example, principle T says that knowledge is true. In conventional possible worlds semantics, that corresponds to the imposition of reflexivity on the model's accessibility relation. Learning semantics also validates T in its standard form, with an explanation—knowledge entails that one has converged to correct belief:

$$\text{T:} \quad \models \mathsf{K}_i\,\phi \to \phi. \tag{33}$$

The rest of the principles on the menu are plainly wrong for cognitively realistic agents. The standard response is to re-interpret K_i vaguely in terms of abilities, obligations, or ideals, but that changes the subject from knowledge to *je ne sais quoi*. It is proposed, instead, to replace material implication $\to$ with conditional feasibility $\langle \mathsf{MD}]\!\!\to_{i,\phi}$. Then, thesis 4 says, plausibly, that there exists a computable inferential procedure that turns knowledge that ϕ into knowledge that one knows that ϕ. The question addressed in this section is which, if any, of the traditional candidate axioms is valid under that interpretation, and under what restrictions, when K_i is interpreted, without equivocation, as inductive knowledge.

14.1. **Deductive Cogency.** Let Δ be a finite set of premises and let Γ be a finite set of conclusions. Suppose that Δ implies Γ, in light of i's information. Maybe i knows neither. But is there any concrete, inferential disposition i could set up in herself to guarantee that if she knows the premises in Δ then she knows the conclusions in Γ as well? Yes, if the premises are inferentially stable, for learning semantics validates the following principle, for finite, disjoint $\Delta, \Gamma \subseteq \mathbf{L}_{\mathsf{BIT}}$ and for arbitrary, finite superset Δ' of Δ that is disjoint from Γ:

$$\text{FD:} \quad \models \; (\mathsf{S}_i\Delta' \;\wedge\; [\,\mathsf{I}\,]_i\,(\Delta \to \Gamma) \;\wedge\; \mathsf{K}_i\Delta) \;\langle \mathsf{MD}]\!\!\to_{i,\Delta'}\; \mathsf{K}_i\Gamma; \tag{34}$$

[26] In the preceding section, it was shown that this timid, non-explanatory strategy is still subject to error.

When $\Delta = \varnothing$ and $\Gamma = \{\phi\}$, thesis (34) collapses to a feasible version of the rule N of necessitation:

(35) $$\text{FN:} \quad \models \; [\mathsf{I}]_i\, \phi \;\; \langle \mathsf{MD}] \!\!\rightarrow_{i,\Delta'} \; \mathsf{K}_i\Gamma.$$

When $\Delta = \{\psi,\ \psi \rightarrow \phi\}$ and $\Gamma = \{\phi\}$, thesis (34) collapses to a feasible version of the standard axiom K:

(36) $$\text{FK:} \quad \models \; (\mathsf{S}_i\Delta' \;\wedge\; \mathsf{K}_i\,\psi \;\wedge\; \mathsf{K}_i(\psi \rightarrow \phi)) \;\; \langle \mathsf{MD}] \!\!\rightarrow_{i,\Delta'} \; \mathsf{K}_i\phi.$$

One may not infer rashly from FN and FK, as one may from the corresponding, traditional axioms N and K, that the knowledge of i is closed under logical consequence, or even that it might be someday. The extension of knowledge by deductive inference must proceed, as it does in the real world, by dint of concrete, cognitive exertion. An inference method that witnesses thesis (34) is *pure deductive inference*—inferring elements of Γ from premises Δ, and for *no other reason*. Then convergence to correct belief that Δ in the actual world results in convergence to true belief that Γ in the actual world and guaranteed, eventual avoidance of error regarding the premises in Δ results in guaranteed, eventual avoidance of error regarding the conclusions in Γ. In that sense, pure deductive inference makes knowledge that Γ *epistemically parasitic* on knowledge that Δ. If the parasitic relationship is disrupted, because i has independent reasons for believing some conclusion $\gamma \in \Gamma$, then i might be disposed to fall into error with respect to γ infinitely often in some possible worlds compatible with current information. The validity of (34) is closely bound to allowance for serendipity. It has already been shown in terms of $\mathsf{G}\,\mathsf{p}_k$ and $\mathsf{GF}\,\mathsf{p}_k$ that (34) fails for learning:

(37) Thesis (34) is invalid with $\tilde{\mathsf{L}}_i, \tilde{\mathsf{Led}}_i, \mathsf{Led}_i$ in place of K_i.

Serendipity raises a cautionary moral about the role of deduction in natural science. The world of science is a "dappled" pastiche of mutually incompatible models and theories and missed connections (Cartwright 1999). Heisenberg and Schrödinger even battled over logically equivalent hypotheses, each of which was rigorously tested over distinct domains of phenomena.[27] When contradictions are found, scientists steer around them until some other experts resolve them, as long as the claims in question remain individually testable. When new logical connections are found between formerly disparate research programs, caution is exercised regarding the drawing of inferences from one program to the other until they are cross-checked by new data. Learning semantics explains that logical conservatism. For suppose that there are two independent research programs studying hypotheses ϕ and ψ, respectively, on the basis of disparate sets of phenomena and then it is discovered by a mathematician that ψ is a deductive consequence of ϕ. What to do? Inferring ψ from ϕ would generate new knowledge that ψ from knowledge that ϕ if inquiry whether ϕ has culminated. But if inquiry whether ψ has culminated in knowledge that $\neg\psi$, then inferring ψ from ϕ would *destroy* knowledge that $\neg\psi$. The contrapositive inference from $\neg\psi$ to $\neg\phi$ is fraught with a similar risk of destroying knowledge that ϕ. Hyper-intensional refusal to fire either inference is guaranteed to preserve knowledge of whichever hypothesis is known

[27]For a version of the history, cf. (van der Werden 1973). Learning semantics allows for the possibility that each scientist knew his own formulation of quantum mechanics at the same time he disputed the competing formulation. Even neighborhood semantics (Scott 1970), which models belief as a set of propositions, cannot model that situation.

and leaves the door open to future empirical evidence to resolve the conflict. So far as inquiry after the truth is concerned, deductive consistency may be a hob-goblin, indeed.

14.2. **Reflection.** Suppose that i knows that ϕ. Evidently, she may fail to know that she knows that ϕ—she may not even conceive of the question whether she knows that ϕ unless she is challenged. Or ϕ may say "i does not believe that she knows me". But inattention and self-referential tricks aside, is i even *capable* of knowing that she knows, even though no bell rings (James 1896) when inductive inquiry succeeds? The prospects look grim:

> ...[Learning in the limit] does not entail that [the learner] knows he knows the answer, since [the learner] may lack any reason to believe that his hypotheses have begun to converge. (Martin and Osherson 1998).

True, i cannot know *infallibly* that she knows some general truth infallibly, because she cannot even know the general truth infallibly. But there is an easy and natural inferential strategy i can adopt to know *inductively* that she knows inductively that ϕ, and so on, to arbitrary iterations. Define iterated knowledge by recursion:

$$\begin{aligned} \mathsf{K}_i{}^0\phi &:= \phi; \\ \mathsf{K}_i{}^{k+1}\phi &:= \mathsf{K}_i\mathsf{K}_i{}^k\phi. \end{aligned}$$

Define the sets of sentences:

$$\begin{aligned} K_i^k(\phi) &= \{\mathsf{K}_i{}^{k'}\phi : k' \leq k\}; \\ K_i^\omega(\phi) &= \bigcup_{k\in\mathbb{N}} K_i^k(\phi). \end{aligned}$$

Then for each finite Δ containing ϕ and disjoint from $K_i^\omega(\phi)$, we have:[28]

$$\text{F4}^*: \quad \models \ (\mathsf{S}_i\Delta \ \wedge \ \mathsf{K}_i\,\phi) \ \langle\mathsf{MD}]\!\!\rightarrow_{i,\Delta} \ K_i^\omega(\phi). \tag{38}$$

As a consequence, we have the following, feasible version of the standard (infeasible) reflection principle 4, for each k:

$$\text{F4}: \quad \models \ (\mathsf{S}_i\Delta \ \wedge \ \mathsf{K}_i\,\phi) \ \langle\mathsf{MD}]\!\!\rightarrow_{i,\Delta} \ \mathsf{K}_i{}^k\,\phi. \tag{39}$$

A simple inferential strategy that witnesses (39) is for i to believe at t that she knew that ϕ at t^* if she never stopped believing that ϕ from t^* until t and to believe that she did not know that ϕ if the alternative case obtains. That inference is intuitive: if i remembers that she retracted ϕ between t^* and the current time t, then the retraction shakes her confidence that she knew that ϕ already at t^*. Otherwise, from i's viewpoint, she had persuasive evidence for ϕ at t^* and nothing in particular has dissuaded her since then, so of course she thinks she knew that ϕ at t^*.

In contrast to the situation for deductive closure, learning that one is learning is easy—learning implies that it is determined that one is learning and whatever is determined can be learned by believing it no matter what and never believing its negation. Having

[28]Strictly speaking, one must restrict $K_i^\omega(\phi)$ to some finite $K_i^k(\phi)$ for the statement to be well-formed, but the proof of validity works for the unrestricted version.

learned whether one has learned whether and having learned that one has learned that are both valid by the same inferential strategy invoked to validate (39). So we have:

(40) Thesis (39) remains valid with $\tilde{\mathsf{K}}_i, \tilde{\mathsf{L}}_i, \tilde{\mathsf{Led}}_i, \mathsf{Led}_i$ in place of K_i.

14.3. **The Unknowable Unknown.** For Plato (1949), the least flattering epistemic condition is *hubris*—failure to know that one does not know. The first step is on the path of inquiry is to eliminate hubris. Thereafter, one comes to know and to know that one knows. But is the fateful, first step feasible? Learning semantics delivers a negative verdict for inductive knowledge, even in the empiricist model $\mathfrak{N}_{t^*}$.

(41) $$\text{F5}': \quad \mathfrak{N}_{t^*} \not\models (\mathsf{S}_i\,\phi \;\wedge\; \neg\mathsf{K}_i\,\phi) \;\; \langle \mathsf{MD}] \!\!\rightarrow_{i,\phi} \;\; \mathsf{K}_i\neg\mathsf{K}_i\,\phi.$$

The convergence required for knowing that one knows parasitically tracks the convergence of knowledge itself. But failure to know inductively may be witnessed only by ugly surprises in the distant future, and the requirement to have converged already to true belief that one will not be surprised in the future occasions the problem of induction, with which we began. For example, suppose that i has seen enough evidence to convince her that $\mathsf{G}\,\mathsf{p}_k$ until such time as some non-k input is received, at which time she would drop her belief that $\mathsf{G}\,\mathsf{p}_k$. Call i's learning method c. Method c yields inductive knowledge that $\mathsf{G}\,\mathsf{p}_k$ in the constantly k world w in which $\mathsf{G}\,\mathsf{p}_k$ is true. Now, suppose that i possesses some magical inferential technique h that guarantees i knowledge now that she does not know that $\mathsf{G}\,\mathsf{p}_k$ if she does not know that $\mathsf{G}\,\mathsf{p}_k$ and that the inferential technique does not alter i's beliefs whether $\mathsf{G}\,\mathsf{p}_k$. Then learning method $h(c)$ must be guaranteed to yield knowledge immediately that c does not produce knowledge that $\mathsf{G}\,\mathsf{p}_k$. Let w_m be the "grue-like" world in which i receives input k until stage m and $k+1$ thereafter. Statement $\mathsf{G}\,\mathsf{p}_k$ is false in w_m, so $h(c)$ stabilizes to belief that $\neg\mathsf{K}_i\mathsf{G}\,\mathsf{p}_k$ immediately in w_m, for *every* m. So $h(c)$ converges to $\neg\mathsf{K}_i\mathsf{G}\,\mathsf{p}_k$ in world w, since w_m agrees empirically with w until m. But, ironically, i *knows* that $\mathsf{G}\,\mathsf{p}_k$ in w because $\mathsf{G}\,\mathsf{p}_k$ is objective in $\mathfrak{N}_{t^*}$ and h holds i's beliefs whether ϕ fixed. So $h(c)$ fails to avoid error in the limit whether $\neg\mathsf{K}_i\mathsf{G}\,\mathsf{p}_k$.

In fact, slight variants of the preceding argument suffice to invalidate the feasible versions of all of the proposed axioms between .4 and 5, so among the standard axioms, only T, FD, and F4 are valid in learning semantics:

$$\begin{array}{llllllll}
\text{FB:} & \mathfrak{N}_{t^*} & \not\models & (\mathsf{S}_i\,\phi \;\wedge & & \neg\phi) & \langle \mathsf{MD}]\!\!\rightarrow_{i,\phi} & \mathsf{K}_i\neg\mathsf{K}_i\,\phi; \\
\text{F.2:} & \mathfrak{N}_{t^*} & \not\models & (\mathsf{S}_i\,\phi \;\wedge & & \neg\mathsf{K}_i\neg\mathsf{K}_i\;\;\neg\phi) & \langle \mathsf{MD}]\!\!\rightarrow_{i,\phi} & \mathsf{K}_i\neg\mathsf{K}_i\,\phi; \\
\text{F.3:} & \mathfrak{N}_{t^*} & \not\models & ((\mathsf{S}_i\,\phi \;\wedge & \mathsf{S}_i\,\psi \;\wedge & \mathsf{K}_i\neg\mathsf{K}_i\;\;\phi) & \langle \mathsf{MD}]\!\!\rightarrow_{i,\phi,\psi} & \mathsf{K}_i\neg\mathsf{K}_i\,\psi) \;\vee \\
 & & \vee & ((\mathsf{S}_i\,\phi \;\wedge & \mathsf{S}_i\,\phi \;\wedge & \mathsf{K}_i\neg\mathsf{K}_i\;\;\psi) & \langle \mathsf{MD}]\!\!\rightarrow_{i,\phi,\psi} & \mathsf{K}_i\neg\mathsf{K}_i\,\phi); \\
\text{F.4:} & \mathfrak{N}_{t^*} & \not\models & (\mathsf{S}_i\,\phi \;\wedge & \neg\phi \;\wedge & \neg\mathsf{K}_i\;\;\neg\phi) & \langle \mathsf{MD}]\!\!\rightarrow_{i,\phi} & \mathsf{K}_i\neg\mathsf{K}_i\phi.
\end{array}$$

It suffices to let $\phi = \mathsf{G}\,\mathsf{p}_k$ and $\psi = \mathsf{G}\,\mathsf{p}_{k'}$, for distinct k, k'.

The same examples refute the corresponding versions of (41-42) for knowing whether, having learned whether, and having learned that:

(42) Theses (41-42) remain invalid with $\tilde{\mathsf{K}}_i, \tilde{\mathsf{Led}}_i, \mathsf{Led}_i$ in place of K_i.

However, it is trivially feasible for i to be learning whether i is not learning whether ϕ when i is not learning whether ϕ—it suffices for i to believe that she is not learning whether ϕ no matter what, since learning begins with operator $[\mathsf{D}]_i$:

$$\text{F5L:} \quad \models (\mathsf{S}_i\,\phi \;\wedge\; \neg\tilde{\mathsf{L}}_i\,\phi)\;\; \langle\mathsf{MD}]\!\!\mapsto_{i,\phi}\;\; \tilde{\mathsf{L}}_i\neg\tilde{\mathsf{L}}_i\,\phi. \tag{43}$$

15. Joint Inductive Knowledge

Plato's original question in the *Meno* (1949) was not what knowledge *is*, but whether virtue can be *taught*. Plato assumed that knowledge can be taught, but when knowledge is inductive, that assumption raises an *epistemological* question. Evidently, a knowledgable expert can *exhibit* her inductive knowledge to her pupils, and on a good day, she might even induce true belief in them, but can she really transfer her inductive *knowledge* to them? In a cooperative epistemic enterprise like education, it is natural to assume that knowledge supervenes *jointly* on the learning strategies of the pupils and of the instructor. In that spirit, this section presents an alternative, *joint* version of learning semantics that is friendlier to cooperative epistemic efforts. In the following section, it is shown how it is jointly feasible for the expert and a room full of pupils to acquire common knowledge of the expert's inductive knowledge.

Let $w \in W$, $\mathbf{c}_{w,t} = (c_{w,1,t}, \ldots c_{w,N,t})$ and $\mathbf{d} \in C^N$. Then let $u[\mathbf{d}/t]$ denote the result of substituting $\mathbf{d}$ for $\mathbf{c}_{w,t}$ in w at t. A *joint* CLM satisfies the following, joint invariance postulate, for each $i \in G$, $w \in W$, $\mathbf{d} \in C^N$, and $t \in T$:

$$s_{i,w}|t \;\;=\;\; s_{i,w[\mathbf{d}/t']}|t. \tag{44}$$

Joint information and determination are defined as follows:

$$\begin{aligned} I_{G,w,t} &= \bigcup_{i \in G} I_{i,w,t}; \\ D_{G,w,t} &= \{u \in I_{G,w,t} : \mathbf{c}_{u,t} = \mathbf{c}_{w,t}\}; \end{aligned}$$

with corresponding operators:

$$\begin{aligned} \|[\mathsf{I}]_G\,\phi\|^t_{\mathfrak{M}_{t^*}} &= \{w \in W : I_{G,w,t} \subseteq \|\phi\|^{t^*}_{\mathfrak{M}_{t^*}}\}; \\ \|[\mathsf{D}]_G\,\phi\|^t_{\mathfrak{M}_{t^*}} &= \{w \in W : D_{G,w,t} \subseteq \|\phi\|^{t^*}_{\mathfrak{M}_{t^*}}\}. \end{aligned}$$

Joint information is weaker than individual information, but joint determination compensates, somewhat, by holding everyone's method fixed. Joint information and determination are no longer guaranteed to be S5 operators, but they *can* be—e.g., everyone gets the same information— so it is useful to have a concise notation for expressing that special case in the object language:

$$\|\mathsf{IS5}_G\|^{t^*}_{\mathfrak{M}_{t^*}} = \{w \in W : (\forall u \in I_{G,w,t^*})\; I_{G,u,t^*} = I_{G,u,t^*}\}.$$

Define joint inductive knowledge for i as before, but with joint determination in place of personal determination:

$$\mathsf{K}_{G,i}\,\phi \;\;:=\;\; \mathsf{GC}_i\,\phi \;\wedge\; [\mathsf{D}]_G\mathsf{FG}\neg\tilde{\mathsf{E}}_i\,\phi.$$

Joint methodological feasibility expresses the existence of a methodological coordination among the agents that brings about ϕ:

$$\|\langle\mathsf{M}\rangle_G\,\phi\|^t_{\mathfrak{M}_{t^*}} = \{w \in W : (\exists \mathbf{d} \in C^N)\; w[\mathbf{d}/t^*] \in \|\phi\|^t_{\mathfrak{M}_{t^*}}\}.$$

To define joint conditional feasibility, let $\mathbf{h} = (h_1, \ldots, h_N)$ be an N-sequence of total recursive functions taking values in C, let $\mathbf{h}(\mathbf{c}) = (h_1(c_1), \ldots, h_N(c_N))$, and let $\boldsymbol{\Delta}$ be an N-sequence of finite subsets of $\mathbf{L}_{\mathsf{BIT}}$. Say that $\mathbf{h}$ *preserves* premises in $\boldsymbol{\Delta}$ if and only if h_i preserves premises in Δ_i, for each $i \in G$ and, similarly, say that $\mathbf{h}$ *depends only* on premises in $\boldsymbol{\Delta}$ if and only if h_i depends only on premises in Δ_i, for each $i \in G$. Then $\mathbf{h}$ is a *joint inference procedure* if and only if $\mathbf{h}$ is an N-sequence of total recursive functions taking values in C that preserves premises in $\boldsymbol{\Delta}$ and that depends only on premises in $\boldsymbol{\Delta}$. Finally, as before, let $\|\psi \ \langle \mathsf{MD}] \!\!\rightarrow_{G,\boldsymbol{\Delta}} \phi\|^t_{\mathfrak{M}_{t^*}}$ denote the set of all $w \in W$ for which there exists joint inference procedure $\mathbf{h}$ such that for all $u \in I_{G,w,t}$:

$$u \in \|\psi\|^t_{\mathfrak{M}_{t^*}} \quad \Rightarrow \quad u[\mathbf{h}/t^*] \in \|\phi\|^t_{\mathfrak{M}_{t^*}}.$$

It remains only to define a joint version of inferential stability. Define $\mathbf{c} \equiv_{\boldsymbol{\Delta}} \mathbf{d}$ to hold if and only if $c_i \equiv_{\Delta_i} d_i$, for all $i \in G$. Let $w \in \|\mathsf{S}_{G,i}\boldsymbol{\Delta}\|^t_{\mathfrak{M}_{t^*}}$ hold if and only if for all $\mathbf{d} \in C^N$ such that $\mathbf{c}_{w,t^*} \equiv_{\boldsymbol{\Delta}} \mathbf{d}$ and for all $u \in D_{G,w,t^*}$, $t \geq t^*$, and $\delta \in \Delta_i$, if we set $u' = u[\mathbf{d}/t^*]$ and $c_i = c_{i,u,t^*} (= c_{i,w,t^*})$ then:

$$\begin{array}{rcl} u \in \|\delta\|^{t^*}_{\mathfrak{M}_{t^*}} & \Leftrightarrow & u' \in \|\delta\|^{t^*}_{\mathfrak{M}_{t^*}}; \\ v_{c_i}(s_{i,u}|t, \delta) & = & v_{d_i}(s_{i,u'}|t, \delta). \end{array}$$

Crucially, a joint version of proposition 2 holds:

Proposition 3. *Suppose that* $\phi \in \Delta_i$ *and* $u \in \|\mathsf{S}_{G,i}\boldsymbol{\Delta}\|^{t^*}_{\mathfrak{M}_{t^*}}$ *and let* $\mathbf{d} \in C^N$ *satisfy* $\mathbf{d} \equiv_{\boldsymbol{\Delta}} \mathbf{c}_{u,t^*}$. *Then:*

$$u \in \|\mathsf{K}_{G,i}\, \phi\|^{t^*}_{\mathfrak{M}_{t^*}} \quad \Rightarrow \quad u[\mathbf{d}/t^*] \in \|\mathsf{K}_{G,i}\, \phi\|^{t^*}_{\mathfrak{M}_{t^*}}. \tag{45}$$

16. Common Inductive Knowledge

Given the joint perspective outlined in the preceding section and some basic assumptions about how the expert and pupils interact, it is jointly feasible for the expert and her pupils to jointly possess the expert's inductive knowledge that ϕ. It suffices that the pupils believe that ϕ if the expert does and suspend belief that ϕ otherwise. Each pupil is then an epistemic parasite of the expert, just as the expert is an epistemic parasite of herself when she infers deductive consequences of what she knows.[29] Educated pupils and news media science reporters can serve, in turn, as experts, resulting in a cascade of joint scientific knowledge through the population—as long as, at the core, some expert has direct inductive knowledge based on experience.[30]

It is a further question whether the pupils and the expert can jointly know that they know, know that they know that they know, etc, all the way to joint, common inductive knowledge that ϕ. Define joint, *mutual, inductive knowledge* to level n as follows:

$$\begin{array}{rcl} {\mathsf{K}_G}^0\, \phi & := & \phi; \\ {\mathsf{K}_G}^{k+1}\, \phi & := & \displaystyle\bigwedge_{i \in G} \mathsf{K}_{G,i} {\mathsf{K}_G}^k\, \phi. \end{array}$$

[29]Indeed, the pupils can know consequences of what the expert knows by deriving them directly from what the expert believes, by the same sort of argument.

[30]More generally, the core expertise is grounded in a research group, but the story with respect to the rest of the population is the same.

Define *common inductive knowledge* that ϕ as the set of sentences:

$$K^{\omega}_G(\phi) \;\;=\;\; \{\mathsf{K}_G{}^k\,\phi : k \in \mathbb{N}\}.$$

It is plausible that a completely trusted, infallible, public announcement that ϕ can generate common knowledge that ϕ. It is less obvious that common *inductive* knowledge is feasible in a room full of computationally bounded pupils who trust their instructor. Learning semantics yields a positive verdict, based on epistemic parasitism and serendipity, in close analogy to the validity argument for F4.

The expert must communicate with the pupils in some way in order to instruct them. It suffices that the pupils receive information sufficient to correctly believe whether the expert believes that ϕ. Let $e \in G$ be the teacher and let $G_- = G \setminus \{e\}$ be the set of pupils. Define the operator "*e teaches* the pupils in G_- whether ϕ" as follows:

$$\mathsf{T}_{G,e}\,\phi \;\;:=\;\; \bigwedge_{j \in G_-} [\,\mathsf{I}\,]_G \mathsf{G}\tilde{\mathsf{C}}_j \mathsf{B}_e\,\phi.$$

Now it is possible to state the *joint feasibility of common inductive knowledge* thesis, which is valid if Δ_e contains ϕ and Δ_i is disjoint from $K^{\omega}_G(\phi)$, for all $i \in G$:

$$\text{(46)} \qquad \text{FC:} \quad \models (\mathsf{IS5}_G \;\wedge\; \mathsf{T}_{G,e}\,\phi \;\wedge\; \mathsf{S}_{G,e}\boldsymbol{\Delta} \;\wedge\; \mathsf{K}_{G,e}\,\phi)\;\langle \mathsf{MD}] \!\!\rightarrow_{G,\boldsymbol{\Delta}}\; K^{\omega}_G(\phi).$$

Although the FC principle concerns common inductive knowledge generated and promulgated by a single expert, it sets the stage for a series of similar results that involve common inductive knowledge generated through the cooperation of a team of experts—a topic of current interest in social epistemology (e.g., Mayo-Wilson 2011).

In dynamic epistemic logic, there are models in which public announcements generate common knowledge of what has been announced (van Benthem 2010). But how do public announcements result in anything more than common knowledge of the fact that the announcement was made? Plausibly, common knowledge of what has been announced is common inductive knowledge grounded in the community's joint strategy to disbelieve sources caught in inconsistencies or lies. One potential extension of FC is to validate the possibility of common inductive knowledge of what is reported in a public announcement in models that allow for false announcements.

A familiar assumption in game theory is that the agents have common knowledge of rationality (Aumann 1995). But how is such knowledge possible and where does it come from? Standard possible worlds semantics has nothing to say, short of a veridical public announcement that all players are rational, but learning semantics provides a plausible, explanatory story. Recall the game-theoretic model described in section 7 above. Violation of the kth level of mutual rationality is detectable by horizontal play in a centipede game of corresponding length. If all of the agents have the disposition to continue playing down at the first move in ever longer centipede games, learning semantics provides a determinate, explanatory, account of how common knowledge of rationality is jointly feasible in such a group. And if every agent is disposed to cooperate by playing sideways for a while, the group can just as easily develop inductive common knowledge of partial cooperation![31]

[31]This application is due to Jennifer Jhun, personal communication.

17. Conclusion and Future Directions

Learning semantics provides a rich, consistent, and workable conceptual framework for modeling interactions between, and iterations of, belief, information, and time, and inductive versions of learning, learnability, having learned, knowing, knowability, and common knowledge. The key feature of the semantics is an assignment of concrete, computational learning methods to each agent at each time. That makes it possible to define inductive learning and knowledge in terms of convergence to the truth and to avoidance of error, on the basis of increasing information through time.

Learning semantics has three important advantages over traditional possible worlds models, for applications involving inductive knowledge and learning. (1) It sidesteps inductive skepticism. (2) It imposes no logical or rational idealizations on the agent's belief states or learning procedures. (3) Its semantic arguments provide concrete, methodological explanations why some principles should be valid and others invalid.

It has been shown that learning semantics validates a cognitively plausible version of the familiar modal system S4 and plausibly refutes all of the standard axioms that have been proposed for epistemic logic beyond S4, when material implication is replaced with conditional feasibility. So the logical sky does not fall, after all, when belief and learning are modeled in a cognitively plausible way. The valid versions of the S4 axioms are explained by epistemic parasitism—the fact that an inferred statement can inherit the convergence conditions essential for knowledge from the convergence conditions possessed by known premises. The invalidity of the remaining axioms is explained by the fact that no inferential procedure can detect immediately that convergence might fail in the future, due to unforeseen surprises. Epistemic parasitism also explains how a knowledgable teacher can *convey* her inductive knowledge to her pupils, as opposed to merely instilling true belief in them, and how inductive common knowledge can spread through a community of passive scientific consumers. Generalization of that idea to inductive learning from the behavior of other learners provides a new understanding of the feasibility of common knowledge of rationality (or of irrational cooperation) in games. Learning semantics explains the scope of learnability in terms of concrete, non-learnability arguments of the sort that are familiar in computational learning theory. It also explains how allowance for serendipity in inductive knowledge both broadens the scope of knowability beyond that of learnability and guarantees that knowability (as opposed to knowledge, itself) is closed under logical consequence. Finally, learning semantics provides a surprising, but plausible, explanation of how one can know one's own Moore sentence $\phi \wedge \neg \mathsf{K}_i \phi$ without ever failing to derive its conjuncts, and without equivocating on the times at which they come to be known. Traditional possible worlds semantics is irrevocably committed to the contrary conclusion, so aloofness from the details of inquiry provides no safe haven from philosophical error.

The explanatory advantages of learning semantics come with a familiar, scientific cost—any formal model of a complex process must abstract, to some extent, from some potentially relevant details. But that is never an argument for giving up on explanation entirely. Instead, one checks whether improvements in the fidelity of one's model result in greater explanatory scope. In that spirit, the paper closes with a tentative discussion of some potential refinements and extensions to the framework developed above. A repeated theme is the importance of greater attention to the epistemic context.

17.1. **Sensitivity and Safety.** Learning semantics was designed to deal with inductive skepticism. It does nothing to avert brain-in-a-vat skepticism—the entire input stream could be the same, whether or not ϕ is true. Relevant alternatives semantics was designed to deal with brain-in-a-vat skepticism, but cannot handle inductive skepticism. Therefore, relevant alternatives semantics and learning semantics are not so much competitors as mutually essential *partners*: the former tosses out virulent but distant possibilities of error that would preclude even convergence to the truth, and the latter eventually weeds out the arbitrarily nearby possibilities of error we couldn't have noticed *yet*. Learning semantics would accommodate the full advantages of both approaches if the key modality D_i were re-interpreted in terms of sensitivity or safety. The change is not entirely trivial, since the "fact" that D_i is an S5 operator is appealed to repeatedly in the preceding development, and each such appeal must be re-examined.

17.2. **Inductive Statistical Knowledge.** Most scientific hypotheses are probabilistic—even the variables of deterministic equations are measured with random error. Such hypotheses can be tested, but a statistical test provides a guaranteed bound on chance of error only when the hypothesis is rejected. So if general statistical hypotheses are knowable, they are knowable only inductively.

A plausible semantics for inductive knowledge of statistical hypothesis ϕ is that i believes that ϕ with high chance that remains high in the actual world and the chance that i believes that ϕ goes to zero if ϕ is false. More ambitiously, one might require, in addition, that the chance that i believes that ϕ converges monotonically to 1 in the actual world. The interpretation of error probabilities requires some temporal gymnastics, as it does in statistical reasoning, itself. Chance is a kind of disposition that governs future events. The fairness of a coin determines chances for sequences of future flips. But the coin might be bent later, after which *different* chances govern sequences of future events—the situation is much the same as it was for learning dispositions. For the chance disposition operative at t, every outcome prior to t has chance 0 or 1, depending on whether it actually occurred. Therefore, the chance that a belief at t^* based on a sample already taken by t^* is 0 or 1 according to the chances operative at t^*. So non-trivial error probabilities must pertain to chances operative at some *reference time* t^{**} prior to sampling—e.g., when the experimental design was originally put into motion. Then the truth of ϕ should also be assessed with respect to the chances operative at t^{**} rather than those operative at t^*.

Since epistemic parasitism pertains to convergence in probability as well as to deterministic convergence, it is anticipated that all of the preceding arguments that depend on epistemic parasitism should generalize to the statistical setting. Also, assuming that successive samples are independent and identically distributed (i.i.d.), successive samples probably provide a better approximation to the fixed, underlying sampling distribution, so the positive results concerning learnability and knowability are also expected to carry over. However, if the sampling distribution may change from time to time, as in time series analysis, extra assumptions are required for convergence in probability to the truth—e.g., that the process under study is periodic, or is driven by hidden states that recur infinitely often (Wei 1989). Analyzing the connection between such assumptions and statistical, inductive knowability is a scientifically relevant, new direction for modal epistemic logic, since statisticians tend to speak of inductive knowledge only informally, if at all.

Aside from its intrinsic interest, the extension of learning semantics to probabilistic theories addresses a puzzle concerning the inductive knowability of future, random outcomes. It is plausible that stochastic theories and models can be known inductively, if inductive knowledge is possible at all. It is far less plausible that random outcomes like coin tosses can be known in advance, even inductively. But the non-statistical version of learning semantics underwrites such knowledge—just make a lucky guess at the outcome (serendipity) and believe the guess until the flip is observed and drop it if it happens to be wrong (Hendricks 2001). The good news is that future coin flips are no longer knowable in statistical learning semantics—the chance of correct belief in the proposition ϕ that the toss will come up heads at future time t is the *joint probability* $p(\mathsf{B}_i\, \phi \,\wedge\, \phi) \leq p(\phi) = 1/2$. What about highly probable future events, such as that your ticket will lose the lottery? They are knowable inductively if their chances of occurring meet the standard for being "high" in the actual convergence condition, but no probabilistic outcome with chance less than one is knowable on the stricter version of the semantics that requires convergence to chance 1 of belief in the actual world.

17.3. **Questions and Coherence.** In light of the aim to model belief more realistically, the logical consistency requirements necessary for knowledge whether ϕ were pared down to the bare minimum required to recover an unambiguous verdict on ϕ for each agent. However, that goes too far. Recall that scientist i can know that the true input sequence is ε by guessing that it is ε until ε is refuted. Suppose that scientist i simultaneously believes *every* hypothesis of the form "the input stream is exactly primitive recursive sequence ε", and is disposed to drop each such hypothesis when it disagrees with the data. Suppose, by serendipity, that the true input stream ε is primitive recursive, so the hypothesis corresponding to ε is true. Then i knows that the future will conform to ε, even though i also believes every possible primitive recursive input stream compatible with current information. That makes inductive knowledge too easy. Furthermore, for someone as aphasic as i, the very concept of belief is called into question. What would i predict to happen at the next stage? Certainly not what she "knows" will happen, since she cannot pick her known theory out of the heap of her alternative, incompatible beliefs. Science may be incoherent overall, but each of its insular paradigms is coherent enough to generate consensus concerning determinate predictions. So normal science within a paradigm is not trivial in the sense under discussion, even though science may remain globally incoherent across paradigms forever. That idea could be modeled in learning semantics by adding a *question under discussion* (q.u.d.) to the epistemic context. The proposal is supported by the current trend in linguistics toward explaining diverse discourse phenomena in terms of such a question (Roberts 2012).

Knowledge of an answer to the q.u.d. requires that the beliefs of the scientist pick out a unique answer, which rules out the easy knowledge just described. The advantages of hyper-intensionality are retained. Inconsistency across question contexts is permitted and even contradictions within a context that do not result in ambiguity concerning the answer selected are still permitted. The correct answer may even be rejected under some logically equivalent formulation, as long as no formulation of any alternative answer is accepted.

17.4. **Feasibility Contextualism.** Epistemic contextualists (e.g., Lewis 1995) hold that the standards for knowledge vary from one context to another—e.g., *raising* a skeptical doubt shifts the epistemic context to one in which the doubt becomes epistemically relevant, so one no longer knows what one knew before the doubt was raised. The idea is appealing, because it does justice both to the plausibility of ordinary knowledge claims and to the apparent force of skeptical doubts. It also addresses a puzzle concerning the psychology of learning. According to learning semantics, it is trivial to know one's own method because the modality D_i holds it fixed and, in the joint version of learning semantics, it is trivial to know what everyone else's method is, because $\mathsf{D}_{G,i}$ holds them all fixed. But when the statement known concerns those very methods, possible worlds involving alternative methods become relevant.

Another plausible, but distinct way in which epistemic standards plausibly depend on context is the intrinsic feasibility of answering the question under discussion. For if "knowledge" is a social encomium whose function is to motivate the overall truth-conduciveness of socially distributed inquiry, then that encomium provides maximum guidance over the full range of epistemic contexts if it it is bestowed only when the agent achieves the best standard of truth-conduciveness achievable with respect to the question in context. Call that natural idea *feasibility contextualism.* For example, concrete, cat-on-the-mat beliefs that can be decided by observation should be, so such knowledge must be safe or sensitive. General laws cannot be known safely or sensitively, but they are learnable, so knowledge should require that they have been learned. More general, untestable theories are unlearnable, but can be known with serendipity, so knowledge with serendipity suffices in that case.

Feasibility contextualism explains why scientists concerned with an inductive question ignore general, philosophical arguments for inductive skepticism, even though they remain fastidious concerning measurement and data analysis. When general theories are at issue, epistemic standards adjust to accommodate knowledge of them, so safety and sensitivity in the short run are no longer required, but error-detection in the limit can still be optimized by catching the errors as soon as possible. Feasibility contextualism also explains why scientists sometimes brand a hypothesis as "metaphysical" if it is difficult to find a plausible, testable articulation of it. In such cases, we simply run out of applicable senses of truth-conduciveness, so skepticism is back on the table.

Feasibility contextualism also helps to resolve a residual puzzle about prediction. It may seem that inductive knowledge, even of future, *deterministic* outcomes is too easy—just guess the outcome and wait to see what happens. But it seems fine—exemplary, even—to deduce the same prediction from an inductively known, universal law. There is a temptation to reach for dark, metaphysical explanations—the law endows the prediction with some ontological "oomph" that a bare prediction lacks. Here is a more concrete, linguistic explanation. When one *infers* a prediction from a law, the law remains in context along with the prediction, and when both the law and the prediction are in context, the operative standard for knowledge is naturally understood to be the strongest standard applicable to *both.* Thus, when the prediction is not inferred from a law, the standard of waiting for sensitivity or safety holds sway, but in light of inferring the prediction from a law, the operative standard is inductive. The idea also explains why the same jarring of intuitions does not accompany the inference of "infinitely often" from "always", for in that case the weaker standard already applies to the conclusion.

17.5. **Justification and Truth-conduciveness.** Scientists prefer unified, cross-testable, explanatory theories over dis-unified, untestable, *ad hoc* theories, a preference popularly known as *Ockham's razor*. Learning semantics, as developed above, does not explain that preference, because a serendipitous guess at a complex law can count as knowledge just as much as a serendipitous guess at a simple one. But the addition of feasibility contextualism suggests such an explanation.[32]

Suppose that the question under discussion is "what is the true form of the polynomial law connecting X and Y?" More precisely, assuming that there exists finite set $S \subseteq \mathbb{N}$ such that the true law has form $Y = f_{\boldsymbol{\theta}}(X) = \sum_{i \in S} \theta_i X^i$, with $\alpha_i \neq 0$ for each $i \in S$, what is S? Assume that the data are arbitrarily small open rectangles in the XY plane guaranteed to intersect the curve $Y = f_{\boldsymbol{\theta}}(X)$.[33] Then there is an important structural relationship between the question and the potential information received by i: any information true of a simpler answer is also compatible with the truth of every more complex answer, whereas some information received if a complex answer is true rules out all simpler and incomparable answers. Instead of viewing those properties as merely symptomatic of the simplicity order, take them as definitive, relative to the question in context.[34] The resulting concept of empirical simplicity assumes alternative guises, depending on the question in context and on the space of possible, future, information states. If one is empirically hunting for new particles or other objects, extra particles make the theory more complex. If one is selecting among theories with free parameters and the parametrization is well-behaved, additional parameters add extra complexity. If one compares theories that entail different symmetry groups, breaking symmetry adds complexity. If one compares theories with more or fewer causes, extra causes add complexity. And so on. It follows from the general definition of empirical simplicity that every learning method capable of inductively learning the true answer to the question can be forced to believe in each successively more complex answer before ultimately converging to the true one. That is an unavoidable, structural feature of the question's semantics, relative to the space of possible information states.

Truth conduciveness is efficient pursuit of the truth. Efficient pursuit entails that one close with the quarry as directly as possible—a random walk or gratuitous aerobatic loops or U-turns during the approach stretch the very concept of pursuit. In the epistemic case, gratuitous loops and U-turns correspond to needless retractions of former beliefs. Thus, retraction minimization is not a mere, pragmatic afterthought—it is constitutive of the

[32] For the details, cf. (Kelly 2010).

[33] In the statistical setting sketched above, the data can be understood, more realistically, as data points sampled independently from the joint distribution generated by the model $Y = f(X) + e$, where e is a normally distributed random variable independent from X and Y that has mean 0 that represents all stray sources of inaccuracy in measurement. Running up the sample size corresponds to narrowing the rectangles in the non-statistical semantics.

[34] That is an over-simplification, but it points in the right direction. Cf. (Kelly 2010) for a better proposal.

very concept of truth-conduciveness. Therefore, feasibility contextualism implies that retractions prior to convergence should be minimized, relative to the current question context. Thus, parties to the question context should forgive methods that change their minds from simpler to more complex theories, since every learning method for the question can be forced to retract that much prior to convergence—but they should forgive no more retractions than those. It can also be shown that the *only* learning methods for the question that minimize worst-case retractions are those that follow Ockham's razor, by selecting the uniquely simplest theory compatible with available information. So Ockham's razor is explained by feasibility contextualism.

The preceding explanation assumes that a fairly rich question is in context, but what if only the known law is in context? Think of the belief $Y = f_{\boldsymbol{\theta}}(X)$ as posing the default, binary question "yes or no" unless a more refined question is in context. There is a learning strategy that retracts at most once when the answer is $Y = f_{\boldsymbol{\theta}}(X)$ (no, yes) and at most twice when the contrary answer is true (yes, no, yes). No tighter bounds are feasible, so that performance is also optimally truth-conducive. The *only* optimal methods are methods that wait for law forms simpler than $Y = f_{\boldsymbol{\theta}}(X)$ to be refuted before yielding a positive verdict for $Y = f_{\boldsymbol{\theta}}(X)$. Thus, feasibility contextualism still entails that $Y = f_{\boldsymbol{\theta}}(X)$ cannot be known unless it is believed in accordance with Ockham's razor.

18. Acknowledgments

This work was supported generously by the John Templeton Foundation, award number 24145. The author has had the basic ideas in mind since the early 1990s and has enjoyed discussions on the topic (in approximately temporal order, and keeping in mind the author's fading memory) with Oliver Schulte, Vincent Hendricks, Stig Andur Pedersen, Nuel Belnap, Johan van Benthem, Horacio Arlo-Costa, Hans Kamp, Fred Dretske, Peter van der Schraaf, Jonathan Vogel, Sherrilyn Roush, Ivan Verano, Hanti Lin, Timothy Williamson, Gregory Wheeler, Alexandru Baltag, Sonja Smets, Gerhard Schurz, Liam Bright, Wesley Holliday, Thomas Icard, Alexandru Radelescu, Clark Glymour, Kevin Zollman, Eric Martin, Konstantin Genin, and Ted Shear, none of whom should be blamed for the outcome. The author is also indebted to Ted Shear and Johan van Benthem for providing detailed comments on drafts. Preliminary versions of the system have been presented in talks at Carnegie Mellon University (2006), the Logic and Methodology Workshop at Stanford (2011), the workshop on Games Interactive Rationality, and Learning in Lund (2012, 2013), and the Workshop on Social Dynamics of Information Change in Amsterdam (2013).

19. Bibliography

Aumann, R. (1995). "Backward Induction and Common Knowledge of Rationality". *Games and Economic Behavior* 8: 619.

Benfey, O. T. (1958). "August Kekulé and the Birth of the Structural Theory of Organic Chemistry in 1848". *Journal of Chemical Education* 35: 21-23.

Carnap, R. (1945). "On Inductive Logic". *Philosophy of Science* 12: 72-97.

Duhem, P. (1914) *La théorie physique son objet et sa structure*, 2nd ed., Paris: Chevalier et Rivire.

Cartwright, N. (1999) "The Dappled World: A Study of the Boundaries of Science". Cambridge: Cambridge University Press.

Church, A. (2009) "Referee Reports on Fitch's 'A Definition of Value'," *New Essays on the Knowability Paradox*. J. Salerno, ed. Oxford: Oxford University Press: 1320.

Dretske, F. (1981) *Knowledge and the Flow of Informatoin*, Cambridge: M.I.T. Press.

Fitch, F. (1963). "A Logical Analysis of some Value Concepts". *The Journal of Symbolic Logic* 28:135-142.

Gierasimczuk, N. (2010) *Knowing One's Limits: Logical Analysis of Inductive Inference*. ILLC Dissertation Series DS-2010-11, Institute for Logic, Language, and Computation.

Glymour, C. (1980) *Theory and Evidence*. Princeton: Princeton University Press.

Gold, E. M. (1967). "Language Identification in the limit". *Information and Control* 10:447-474.

Hempel, C. (1945) "Studies in the Logic of Confirmation". *Mind* 213: 1-26.

Hendricks, V. (2001). *The Convergence of Scientific Knowledge*. Dordrecht: Springer.

Holliday, W. (2012). "Epistemic Closure and Epistemic Logic I: Relevant Alternatives and Subjunctivism". *Journal of Philosophical Logic*, forthcoming.

Jain, S., D. Osherson, J. Royer, A. Sharma. (1999) *Systems that Learn: An Introduction to Learning Theory*, 2nd ed. New York: Bradford.

James, W. (1896). "The Will to Believe". *The New World*:

Kamp, H. (1971). "Formal Properties of 'Now"'. *Theoria* 27:227-274.

Kant, I. (1782/1787). *Critique of Pure Reason*. P. Guyer and A. Wood, trans. Cambridge: Cambridge University Press. 1997.

Kelly, K. T. (1996). *The Logic of Reliable Inquiry*. New York: Oxford University Press.

Kelly, K. T. (2001). "The Logic of Success". *British Journal for the Philosophy of Science, special millennium issue* 51: 639-666.

Kelly, K. T. (2010) "Simplicity, Truth, and Probability", *Handbook for the Philosophy of Statistics*. Prasanta S. Bandyopadhyay and Malcolm Forster eds., Dordrecht: Elsevier.

Klein, P., and Warfield, T. A., (1994) "What Price Coherence?," *Analysis* 54: 129132.

Laudan, L. (1981). "A Confutation of Convergent Realism". *Philosophy of Science* 48: 19-49.

Lewis, D. (1996). "Elusive Knowledge". *Australasian Journal of Philosophy*, 74: ***-**.

Martin, E. and D. Osherson (1998) "Elements of Scientific Inquiry". Cambridge: M.I.T. Press.

Mayo-Wilson, C. (2011). "The Problem of Piece-meal Induction". *Philosophy of Science* 78: 864-874.

Nozick, R. (1981). *Philosophical Explanations.* Cambridge: Harvard University Press.

Peirce, C. S. (1878). "How to Make our Ideas Clear". *Popular Science Monthly.* 12: 286-302.

Poincare, H. (1904). *Science and Hypothesis.* New York: The Walter Scott Publishing Co.

Plato (1949). *Meno.* B. Jowett, trans. Indianapolis: Bobbs-Merrill.

Popper, K. R. (1935). *Logik der Forschung.* Vienna: Springer.

Putnam, H. (1963). "'Degree of Confirmation' and Inductive Logic". *The Philosophy of Rudolf Carnap*, ed. A. Schilpp. LaSalle: Open Court.

Pritchard, D. (2007). "Anti-Luck Epistemology". *Synthese* 158: 277-98.

Reichenbach, H. (1949). *The Theory of Probability.* London: Cambridge University Press.

Roberts, C. (2012) "Information Structure: Towards an integrated formal theory of pragmatics", *Semantics and Pragmatics*, 2012.

Roush, S. (2007). *Tracking Truth: Knowledge, Evidence, and Science.* Oxford: Oxford University Press.

Roush, S. (2010). "Closure on Skepticism". *Journal of Philosophy* 107: 243-256.

Scott, D. (1970). "Advice in Modal Logic". *Philosophicawl Problems in Logic*, ed. K. Lambert: 143-173.

Sosa, E. (1999). "How to Defeat Opposition to Moore". *Philosophical Perspectives* 13: 141-54.

Sosa, E., (1980). "The Raft and the Pyramid: Coherence versus Foundations in the Theory of Knowledge," *Midwest Studies in Philosophy* 5: 325.

Vogel, J. (1987). "Tracking, Closure, and Inductive Knowledge". *The Possibility of Knowledge: Nozick and His Critics.* S. Luper-Foy, ed. Totowa: Rowman and Littlefield.

van Benthem, J. (2011) *Logical Dynamics of Information and Interaction*, Cambridge: Cambridge University Press.

van Benthem, J. (2010) *Modal Logic for Open Minds*, Standord: CSLI Lecture Notes.

van der Werden, (1973) "From Matrix Mechanics and Wave Mechanics to Unified Quantum Mechanics". *The Physicist's Conception of Nature*, J. Mehra, ed. Dordrecht: Reidel. pp. 276-293.

Yap, A. and Tomohiro, H. (2009) "Dynamic Epistemic Logic and Branching Temporal Structure", *Synthese* 169: 259-281.

Williamson, T. (1993). "Verificationism and Non-Distributive Knowledge". *Australasian Journal of Philosophy* 71: 7886.

Wei, W. (1989). *Time series analysis: Univariate and multivariate methods.* New York: Addison-Wesley.

Williamson, T. (2000). *Knowledge and its Limits.* Oxford: Oxford University Press.

20. Proofs of Propositions

Proof of proposition 1. Just let $L_d(\sigma, \psi)$ return 1 if $\psi = @_{t^*}\,\phi$ and σ is an initial segment of $s_{i,u,t}$ and return 0 otherwise. □

Proof of proposition 2. Abbreviate:

$$\begin{aligned} c &= c_{i,u,t^*}; \\ x &= u[d/i, t^*]. \end{aligned}$$

Assume that $\phi \in \Delta$ and that:

(47) $$d \equiv_\Delta c;$$

(48) $$u \in \|\mathsf{S}_i \Delta\|^{t^*}_{\mathfrak{M}_{t^*}};$$

(49) $$u \in \|\mathsf{K}_i\, \phi\|^{t^*}_{\mathfrak{M}_{t^*}}.$$

From (49) we have:

(50) $$u \in \|\mathsf{GC}_i\, \phi\|^{t^*}_{\mathfrak{M}_{t^*}};$$

(51) $$y \in \|\mathsf{FG}\neg\tilde{\mathsf{E}}_i\, \phi\|^{t^*}_{\mathfrak{M}_{t^*}}, \text{ for all } y \in D_{i,u,t^*}.$$

It suffices to show that:

(52) $$x \in \|\mathsf{GC}_i\, \phi\|^{t^*}_{\mathfrak{M}_{t^*}};$$

(53) $$y \in \|\mathsf{FG}\neg\tilde{\mathsf{E}}_i\, \phi\|^{t^*}_{\mathfrak{M}_{t^*}}, \text{ for all } y \in D_{i,x,t^*}.$$

From (47-48), we have that:

(54) $$u \in \|\phi\|^{t^*}_{\mathfrak{M}_{t^*}} \Leftrightarrow x \in \|\phi\|^{t^*}_{\mathfrak{M}_{t^*}};$$

(55) $$u \in \|\mathsf{G}[\mathsf{B}]_i\, \phi\|^{t^*}_{\mathfrak{M}_{t^*}} \Leftrightarrow x \in \|\mathsf{G}[\mathsf{B}]_i\, \phi\|^{t^*}_{\mathfrak{M}_{t^*}};$$

(56) $$u \in \|\mathsf{G}\langle\mathsf{B}\rangle_i\, \phi\|^{t^*}_{\mathfrak{M}_{t^*}} \Leftrightarrow x \in \|\mathsf{G}\langle\mathsf{B}\rangle_i\, \phi\|^{t^*}_{\mathfrak{M}_{t^*}}.$$

So requirement (52) follows from (50).

For requirement (53), let $y \in D_{i,x,t^*}$. Then $s_{i,y}|t^* = s_{i,x}|t^* = s_{i,u[d/i,t^*]}|t^*$. So $s_{i,y}|t^* = s_{i,u}|t^*$, by (1). Let $z = y[c/i, t^*]$. So $s_{i,z}|t^* = s_{i,u}|t^*$, again by (1) and, hence, $z \in D_{i,u,t^*}$. So it follows from (51) that:

(57) $$z \in \|\mathsf{FG}\neg\tilde{\mathsf{E}}_i\, \phi\|^{t^*}_{\mathfrak{M}_{t^*}};$$

and from (47-48) that:

(58) $$y \in \|\phi\|^{t^*}_{\mathfrak{M}_{t^*}} \Leftrightarrow z \in \|\phi\|^{t^*}_{\mathfrak{M}_{t^*}};$$

(59) $$y \in \|\mathsf{FG}[\mathsf{B}]_i\, \phi\|^{t^*}_{\mathfrak{M}_{t^*}} \Leftrightarrow z \in \|\mathsf{FG}[\mathsf{B}]_i\, \phi\|^{t^*}_{\mathfrak{M}_{t^*}};$$

(60) $$y \in \|\mathsf{FG}\langle\mathsf{B}\rangle_i\, \phi\|^{t^*}_{\mathfrak{M}_{t^*}} \Leftrightarrow z \in \|\mathsf{FG}\langle\mathsf{B}\rangle_i\, \phi\|^{t^*}_{\mathfrak{M}_{t^*}}.$$

Requirement (53) follows directly from (57-60). □

Proof of proposition 3. Let $\mathbf{d} \in C^N$ and let $u \in W$. Abbreviate:

$$\begin{aligned} \mathbf{c} &= \mathbf{c}_{i,u,t^*}; \\ x &= u[\mathbf{d}/t^*]. \end{aligned}$$

Assume that $\phi \in \Delta_i$ and that:

$$d_i \equiv_\phi c_i; \tag{61}$$

$$u \in \|\mathsf{S}_{G,i}\boldsymbol{\Delta}\|^{t^*}_{\mathfrak{M}_{t^*}}; \tag{62}$$

$$u \in \|\mathsf{K}_{G,i}\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}. \tag{63}$$

Proceed as in the preceding proof, with D_{G,u,t^*}, D_{G,x,t^*} in place of D_{i,u,t^*}, D_{i,x,t^*}. The argument for requirement (52) is the same as before. For requirement (53), let $y \in D_{G,x,t^*}$. So $y \in D_{i,x,t^*}$, for some $i \in G$. Then $s_{i,y}|t^* = s_{i,x}|t^* = s_{i,u[\mathbf{d}/t^*]}|t^*$. So $s_{i,y}|t^* = s_{i,u}|t^*$, by (44). Let $z = y[\mathbf{c}/t^*]$. So $s_{i,z}|t^* = s_{i,u}|t^*$, again by (44) and, hence, $z \in D_{i,u,t^*} \subseteq D_{G,u,t^*}$. Continue as in the preceding proof. □

21. Proofs of Selected Statements

Proof of (7) *and* (8). Let $w \in W$ be given. To witness the first claim, define learning method c so that:

$$L_c(\sigma, \phi) = \begin{cases} 1 & \text{if } \phi = @_{t^*}\mathsf{G}\,\mathsf{p}_k \text{ and } (\forall t : t^* \leq t \leq \mathsf{lh}(\sigma))\ \sigma(t) = k; \\ 1 & \text{if } \phi = @_{t^*}\neg\mathsf{G}\,\mathsf{p}_k \text{ and } (\exists t : t^* \leq t \leq \mathsf{lh}(\sigma))\ \sigma(t) \neq k; \\ 0 & \text{otherwise.} \end{cases}$$

The method that witnesses the second claim is similar, except that $\neg$ and $\neq$ are moved from the second clause to the first. □

Proof of (9) *and* (10). The proof of the second statement is similar to that of the first. For the first statement, suppose for contradiction that c satisfies (3) and (4). It suffices to construct $\varepsilon \in E_0$ such that (*) both (3) and (4) are false in arbitrary world w such that $e_w = \varepsilon$. A purely learning theoretic argument suffices. Construct ε by adding chunks in successive stages as follows, where $c = h(c_{w',i,t^*})$. At stage 0, present σ. Let $n > 0$. At stage $2n$, present k until L_c returns 1 for $@_{t^*}\mathsf{FG}\,\mathsf{p}_k$. Learning function L_c must return 1 for $@_{t^*}\mathsf{FG}\,\mathsf{p}_k$ eventually, because if L_c never takes the bait, you continue to present k and L_c fails to converge to belief that $@_{t^*}\mathsf{FG}\,\mathsf{p}_k$ even though it is true, contradicting the hypothesis. At that point, proceed to stage $2n+1$. At stage $2n+1$, the demon presents $k+1$ until L_c returns 0 for $@_{t^*}\mathsf{FG}\,\mathsf{p}_k$. Learning function L_c must return 0 for $@_{t^*}\mathsf{FG}\,\mathsf{p}_k$ eventually, because if L_c never takes the bait, you continue to present $k+1$ and L_c fails to converge to belief that $@_{t^*}\neg\mathsf{G}\,\mathsf{p}_k$ even though it is true, contradicting the hypothesis. At that point, proceed to stage $2n+2$. You pass through each stage, producing ε that satisfies (*). □

Proof of (29). The proof follows (Kelly 1996, proposition 7.15). Suppose the contrary. Then we can use the witnessing L_d and $u \in I_{i,w,t^*}$ to compute $g(t)$, for $t \geq t^*$ (for $t < t^*$, use a lookup table). Say that finite input sequence σ of length t is t'-dead if and only if $L_d(\sigma', @_{t^*}\,\phi) = 0$, for each extension σ' of σ of length t'. By (24), $g|(t+1)$ is never t'-dead, but by König's lemma and (27), there exists $t' \geq t+1$ such that every σ of length $t+1$

that is distinct from $g|t$ is t'-dead. Then $g|(t+1)$ is the unique sequence σ that is not t'-dead. Return the last entry of that sequence. $\square$

Proof of (32). By hypothesis, ϕ is knowable in w at t^*. Since ϕ is knowable, let L_c and world $u \in I_{i,w,t^*}$ witness that fact. Let L_d believe that ϕ in all circumstances and believe, deny, or suspend belief for both $\neg \mathsf{K}_i\, \phi$ and $\mathsf{Mo}_i\, \phi$ whenever L_c does the same for ϕ. Since ϕ is assumed to be false in some world compatible with information, i does not know that ϕ. Recall that in $\mathfrak{N}_{t^*}$, (i) the inputs to i do not depend on i's learning method and (ii) the truth value of ϕ does not depend on i's learning method. Due to L_d's dogmatic belief that ϕ, the case hypothesis, and (i) and (ii), there is no world in I_{i,w,t^*} in which $\mathsf{K}_i\phi$ is true, so we have that $[\,\mathsf{I}\,]_i(\mathsf{Mo}_i\, \phi \leftrightarrow \phi)$ is true in w. So by (i) and (ii), agent i knows that $\mathsf{Mo}_i\, \phi$. By construction, i is conjunctively cogent with respect to $\mathsf{Mo}_i\, \phi$. $\square$

Proof of (34). Let Δ, Γ be finite and mutually disjoint subsets of $\mathbf{L}_{\mathsf{BIT}}$. Let $\Delta \subseteq \Delta'$ and $\Delta' \cap \Gamma = \varnothing$. Define total recursive g such that:

$$g(c, \langle\sigma\rangle, \ulcorner\phi\urcorner)) = \begin{cases} 1 & \text{if } \phi = @_{t^*}\gamma \ \wedge\ \gamma \in \Gamma \ \wedge\ (\forall \delta \in \Delta)\ L_c(\sigma, @_{t^*}\delta) = 1; \\ 0 & \text{if } \phi = @_{t^*}\neg\gamma \ \wedge\ \gamma \in \Gamma; \\ L_c(\sigma, \phi) & \text{otherwise.} \end{cases}$$

The following lemma is a familiar consequence of the s-m-n theorem of recursive function theory:

(64) $$(\forall \text{ t.r. } f)(\exists \text{ t.r. } h)(\forall c, x, y \in \mathbb{N})\ \phi_{h(c)}(x, y) = f(c, x, y).$$

Apply (64) to obtain total recursive h such that $L_{h(c)}(\sigma, \phi) = g(c, \langle\sigma\rangle, \ulcorner\phi\urcorner)$. By the definition of h and the fact that Δ' is disjoint from Γ, we have that:

(65) $$c \ \equiv_{\Delta'} \ h(c),$$

for each $c \in C$, and that for all $z \in W$, $t \in T$ and $\gamma \in \Gamma$:

(66) $$L_{h(c)}(s_{i,z}|t, @_{t^*}\neg{\mathsf{K}_i}^k\, \gamma) = 0;$$

(67) $$L_{h(c)}(s_{i,z}|t,\ @_{t^*}{\mathsf{K}_i}^k\, \gamma) = 1 \ \Leftrightarrow\ (\forall \delta \in \Delta)\, L_{h(c)}(s_{i,z}|t', @_{t^*}\, \delta) = 1.$$

Suppose that $u \in I_{i,w,t^*}$ satisfies:

(68) $$u \ \in \ \|\mathsf{S}_i\Delta'\|^{t^*}_{\mathfrak{M}_{t^*}};$$

(69) $$u \ \in \ \|[\,\mathsf{I}\,]_i(\Delta \to \Gamma)\|^{t^*}_{\mathfrak{M}_{t^*}};$$

(70) $$u \ \in \ \|\mathsf{K}_i\Delta\|^{t^*}_{\mathfrak{M}_{t^*}}.$$

Abbreviate:

$$\begin{aligned} c &= c_{i,u,t^*}; \\ x &= u[h(c)/i, t^*]. \end{aligned}$$

So from (65),(68) and (70), obtain via proposition 2 that for each $\delta \in \Delta$:

(71) $$x \ \in \ \|\mathsf{K}_i\, \delta\|^{t^*}_{\mathfrak{M}_{t^*}}.$$

So for each $\delta \in \Delta$:

$$x \in \|\mathsf{GC}_i\,\delta\|^{t^*}_{\mathfrak{M}_{t^*}}; \tag{72}$$

$$y \in \|\mathsf{FG}\neg\tilde{\mathsf{E}}_i\,\delta\|^{t^*}_{\mathfrak{M}_{t^*}},\ \text{for all } y \in D_{i,x,t^*}. \tag{73}$$

It suffices to show the following requirements, for each $\gamma \in \Gamma$:

$$x \in \|\mathsf{GC}_i\,\gamma\|^{t^*}_{\mathfrak{M}_{t^*}}; \tag{74}$$

$$y \in \|\mathsf{FG}\neg\tilde{\mathsf{E}}_i\,\gamma\|^{t^*}_{\mathfrak{M}_{t^*}},\ \text{for all } y \in D_{i,x,t^*}. \tag{75}$$

Let $\gamma \in \Gamma$. For requirement (74), we have by (1) that $x \in I_{i,u,t^*}$, so (69) and (72) yield that:

$$x \in \|\gamma\|^{t^*}_{\mathfrak{M}_{t^*}}. \tag{76}$$

So (72) and (76), together with properties (66-67), yield requirement (74). For requirement (75), suppose that $y \in D_{i,x,t^*}$. So by (1), $y \in I_{i,u,t^*}$. So (*) together with (69) and (73) yield requirement (75). $\square$

Proof of statement (39). Define total recursive f as follows:

$$f(c, \langle\sigma\rangle, \mathsf{g}(\psi)) = \begin{cases} 1 & \text{if } (\exists k)\ \psi = @_{t^*}\mathsf{K}_i{}^k\phi\ \wedge \\ & \quad (\forall t' : t^* \le t' \le t)(\psi = @_{t^*}\mathsf{K}_i{}^k\phi\ \wedge\ L_c(\sigma|t', \phi) = 1); \\ 0 & \text{if } (\exists k)\ \psi = @_{t^*}\mathsf{K}_i{}^k\phi\ \wedge \\ & \quad (\exists t' : t^* \le t' \le t)(\psi = @_{t^*}\mathsf{K}_i{}^k\phi\ \wedge\ L_c(\sigma|t', \phi) = 0); \\ 0 & \text{if } (\exists k)\ \psi = @_{t^*}\neg\mathsf{K}_i{}^k\phi\ \wedge \\ & \quad (\forall t' : t^* \le t' \le t)(\psi = @_{t^*}\mathsf{K}_i{}^k\phi\ \wedge\ L_c(\sigma|t', \phi) = 1); \\ 1 & \text{if } (\exists k)\ \psi = @_{t^*}\neg\mathsf{K}_i{}^k\phi\ \wedge \\ & \quad (\exists t' : t^* \le t' \le t)(\psi = @_{t^*}\mathsf{K}_i{}^k\phi\ \wedge\ L_c(\sigma|t', \phi) = 0); \\ L_c(\sigma, \phi) & \text{otherwise.} \end{cases}$$

Apply (64) to obtain h such that $L_{h(c)}(\sigma, \psi) = f(c, \langle\sigma\rangle, \ulcorner\phi\urcorner)$, for all $c \in \mathbb{N}$. Suppose that Δ includes ϕ and is disjoint from $K_i^\omega(\phi)$. By the definition of h, we have that for all $c \in C$:

$$c \equiv_\Delta h(c); \tag{77}$$

so h preserves Δ. Moreover, by construction, h depends only on Δ. Furthermore, for all $z \in W$, $t \in T$, and $k \in \mathbb{N}$:

$$L_{h(c)}(s_{i,z}|t, @_{t^*}\mathsf{K}_i{}^k\,\phi) = 1 \ \Leftrightarrow\ (\forall t' : t^* \le t' \le t)\, L_{h(c)}(s_{i,z}|t', @_{t^*}\,\phi) = 1; \tag{78}$$

$$L_{h(c)}(s_{i,z}|t, @_{t^*}\neg\mathsf{K}_i{}^k\,\phi) = 1 \ \Leftrightarrow\ (\exists t' : t^* \le t' \le t)\, L_{h(c)}(s_{i,z}|t', @_{t^*}\,\phi) = 0. \tag{79}$$

Suppose that $u \in I_{i,w,t^*}$ satisfies:

$$u \in \|\mathsf{S}_i\Delta\|^{t^*}_{\mathfrak{M}_{t^*}}; \tag{80}$$

$$u \in \|\mathsf{K}_i\phi\|^{t^*}_{\mathfrak{M}_{t^*}}. \tag{81}$$

Abbreviate:

$$\begin{aligned} c &= c_{i,u,t^*}; \\ x &= u[h(c)/i, t^*]. \end{aligned}$$

From (77), (80) and (81), obtain via proposition 2 that $x \in \|\mathsf{K}_i\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}} = \|\mathsf{K}_i{}^1\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}$. Therefore, $x \in \|\phi\|^{t^*}_{\mathfrak{M}_{t^*}} = \|\mathsf{K}_i{}^0\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}$. So we have the base case $x \in \|K^1(\phi)\|^{t^*}_{\mathfrak{M}_{t^*}}$.

Next, assume for induction that $x \in \|K^{k+1}(\phi)\|^{t^*}_{\mathfrak{M}_{t^*}}$. So:

$$x \in \|\mathsf{K}_i\mathsf{K}_i{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}; \tag{82}$$

and, therefore:

$$x \in \|\mathsf{G}\mathsf{C}_i\mathsf{K}_i{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}; \tag{83}$$

$$y \in \|\mathsf{F}\mathsf{G}\neg\tilde{\mathsf{E}}\mathsf{K}_i{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}, \text{ for all } y \in D_{i,x,t^*}\ . \tag{84}$$

For $x \in \|K^{k+2}(\phi)\|^{t^*}_{\mathfrak{M}_{t^*}}$, it suffices to show that: $x \in \|\mathsf{K}_i\mathsf{K}_i\mathsf{K}_i{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}$. For that, it suffices, in turn, to show:

$$x \in \|\mathsf{G}\mathsf{C}_i\mathsf{K}_i\mathsf{K}_i{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}; \tag{85}$$

$$x \in \|\mathsf{F}\mathsf{G}\neg\tilde{\mathsf{E}}\mathsf{K}_i\mathsf{K}_i{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}, \text{ for all } y \in D_{i,x,t^*}\ . \tag{86}$$

Requirement (85) expands to the requirements:

$$x \in \|\mathsf{K}_i\mathsf{K}_i{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}; \tag{87}$$

$$x \in \|\mathsf{G}[\mathsf{B}]_i\mathsf{K}_i\mathsf{K}_i{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}; \tag{88}$$

$$x \in \|\mathsf{G}\langle\mathsf{B}\rangle_i\mathsf{K}_i\mathsf{K}_i{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}. \tag{89}$$

Requirement (87) is just (82). Hence, (83) yields:

$$x \in \|\mathsf{G}[\mathsf{B}]_i\mathsf{K}_i{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}; \tag{90}$$

$$x \in \|\mathsf{G}\langle\mathsf{B}\rangle_i\mathsf{K}_i{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}. \tag{91}$$

Requirements (88-89) follow from (90-91) and properties (78-79) of h.

For requirement (86), suppose that $y \in D_{i,x,t^*}$. It suffices to show that for all $y \in D_{i,x,t^*}$:

$$y \in \|\mathsf{G}\mathsf{F}[\mathsf{B}]_i\neg\mathsf{K}_i\mathsf{K}_i{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}} \;\Rightarrow\; y \notin \|\mathsf{K}_i\mathsf{K}_i{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}; \tag{92}$$

$$y \in \|\mathsf{G}\mathsf{F}[\mathsf{B}]_i\mathsf{K}_i\mathsf{K}_i{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}} \;\Rightarrow\; y \in \|\mathsf{K}_i\mathsf{K}_i{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}. \tag{93}$$

(94)

For requirement (92), suppose that:

$$y \in \|\mathsf{G}\mathsf{F}[\mathsf{B}]_i\neg\mathsf{K}_i\mathsf{K}_i{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}} \tag{95}$$

Then by property (79) of h, there exists $t \geq t^*$ such that $y \notin \|\mathsf{B}_i\,\phi\|^{t}_{\mathfrak{M}_{t^*}}$, so by property (78), we have that $y \notin \|\mathsf{B}_i\mathsf{K}_i{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}$. So $y \notin \|\mathsf{K}_i\mathsf{K}_i{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}$.

For requirement (93), suppose that:

$$y \in \|\mathsf{G}\mathsf{F}[\mathsf{B}]_i\mathsf{K}_i\mathsf{K}_i{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}} \tag{96}$$

For the consequent $y \in \|\mathsf{K}_i\mathsf{K}_i^{\;k}\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}$, it suffices, as usual, to show the requirements:

$$y \;\in\; \|\mathsf{GC}_i\mathsf{K}_i^{\;k}\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}; \tag{97}$$

$$z \;\in\; \|\mathsf{FG}\neg\tilde{\mathsf{E}}\mathsf{K}_i^{\;k}\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}, \text{ for all } z \in D_{i,y,t^*}\;. \tag{98}$$

Requirement (98) is just (84), since $D_{i,y,t^*} = D_{i,u,t^*}$. Requirement (97) expands to:

$$y \;\in\; \|\mathsf{K}_i^{\;k}\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}; \tag{99}$$

$$y \;\in\; \|\mathsf{G}[\mathsf{B}]_i\mathsf{K}_i^{\;k}\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}; \tag{100}$$

$$y \;\in\; \|\mathsf{G}\langle\mathsf{B}\rangle_i\mathsf{K}_i^{\;k}\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}. \tag{101}$$

For requirement (99), we have from (96) and property (78) of h that $y \in \|\mathsf{GF}[\mathsf{B}]_i\mathsf{K}_i^{\;k}\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}$. So $y \in \|\mathsf{K}_i^{\;k}\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}$, by (84). For requirement (100), note that (96), along with property (78) of h implies that $y \in \|\mathsf{G}[\mathsf{B}]_i\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}$, which implies requirement (100) in light of property (78) and requirement (101) in light of property (79). □

Proof of statement (40). For the Led_i case, follow the proof of (39) with Led_i in place of K_i and $\tilde{\mathsf{C}}_i$ in place of $\tilde{\mathsf{E}}_i$. For the $\tilde{\mathsf{L}}_i$ case, make corresponding substitutions and ignore the actual convergence requirements. For the $\tilde{\mathsf{Led}}_i$ case, add cases for actual convergence to true belief that $\neg\phi$. For the $\tilde{\mathsf{K}}_i$ case, do the same, but retain $\tilde{\mathsf{C}}_i$ in place of $\tilde{\mathsf{E}}_i$. □

Proof of statement (41). Let $w = (\varepsilon, \mathbf{c})$ be a world in $\mathfrak{N}_{t^*}$. Let total recursive h preserve belief whether $\phi = \mathsf{G}\,\mathsf{p}_k$. Let c^* be as in the proof of statement (7). Let $\mathbf{c} \in C^N$ and let $w_{\varepsilon'} = (\varepsilon', \mathbf{c}[c^*/i]_{t^*})$, for arbitrary $\varepsilon' \in E_0$. Let $\tau(t) = \varepsilon(t)$ for $t < t^*$ and let $\tau(t) = k$ for $t \geq t^*$. Let $\tau_t(t') = \tau(t')$ for $t' \geq t$ and let $\tau_t(t') = k+1$ for $t' \geq t$. It is easy to verify that for all $t \geq t^*$:

$$w_\tau \;\in\; \|\mathsf{K}_i\mathsf{G}\,\mathsf{p}_k\|^{t^*}_{\mathfrak{N}_{t^*}}; \tag{102}$$

$$w_{\tau_t} \;\in\; \|\neg\mathsf{K}_i\mathsf{G}\,\mathsf{p}_k\|^{t^*}_{\mathfrak{N}_{t^*}}. \tag{103}$$

Since the truth of $\mathsf{G}\,\mathsf{p}_k$ does not depend on methods in $\mathfrak{N}_{t^*}$, we have for all $t \geq t^*$ that:

$$w_{\tau_t} \;\in\; I_{i,w,t^*} \cap \|\mathsf{S}_i\,\mathsf{G}\,\mathsf{p}_k\|^{t^*}_{\mathfrak{N}_{t^*}} \cap \|\neg\mathsf{K}_i\mathsf{G}\,\mathsf{p}_k\|^{t^*}_{\mathfrak{N}_{t^*}}. \tag{104}$$

So it suffices to show that $w_{\tau_t}[h(c^*)/i, t^*] \notin \|\mathsf{K}_i\neg\mathsf{K}_i\mathsf{G}\,\mathsf{p}_k\|^{t^*}_{\mathfrak{N}_{t^*}}$. For that it suffices to show that at least one of the following statements holds:

$$w_{\tau_t}[h(c^*)/i, t^*] \notin \|\mathsf{GC}_i\neg\mathsf{K}_i\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}; \tag{105}$$

$$w_{\tau_t}[h(c^*)/i, t^*] \notin \|[\mathsf{D}]_i\mathsf{FG}\neg\tilde{\mathsf{E}}_i\neg\mathsf{K}_i\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}. \tag{106}$$

Case 1: $w_{\tau_t}[h(c^*)/i, t^*] \notin \|\mathsf{GB}_i\neg\mathsf{K}_i\mathsf{G}\,\mathsf{p}_k\|^{t^*}_{\mathfrak{N}_{t^*}}$, for some $t \geq t^*$. So (105) holds, in light of (103).

Case 2: $w_{\tau_t}[h(c^*)/i, t^*] \in \|\mathsf{GB}_i\neg\mathsf{K}_i\mathsf{G}\,\mathsf{p}_k\|^{t^*}_{\mathfrak{N}_{t^*}}$, for all $t \geq t^*$. Then since $\tau|t = \tau_t|t$, for each $t \geq t^*$, we have that $w_\tau[h(c^*)/i]_{\geq t^*} \in \|\mathsf{GB}_i\neg\mathsf{K}_i\mathsf{G}\,\mathsf{p}_k\|^{t^*}_{\mathfrak{N}_{t^*}}$. Note that $w_\tau \in I_{w_{\tau_t},i,t^*}$ by construction and (1). So (106) holds, in light of (102). □

Proof of statements (42-42). One merely has to check that the respective antecedents of the various conditionals are satisfied by each world w_{τ_t} in the proof of (41). For (42), observe that $\neg\phi$ is true in w_{τ_t}, by construction. For (42), observe that c^* suspends belief concerning $\neg\mathsf{K}_i\neg\phi$. For (42), observe both that c^* suspends belief concerning $\neg\phi$ and that $\neg\phi$ is true in w_{τ_t}. For (42), let $w \in W$ and let total recursive h preserve both ϕ and ψ. To refute the second disjunct of (42) in w, let c^{**} follow the strategy of c^* with respect to ϕ, except that c^{**} believes that $\neg\mathsf{K}_i\,\psi$ no matter what. Then, due to c^{**}'s suspension of belief whether ψ at t^*, we have that c^{**} witnesses the truth of $\mathsf{K}_i\neg\mathsf{K}_i\,\psi$ in every world, so the argument for (41) establishes the falsehood of the second disjunct of (42) in w. Reversing the roles of ϕ and ψ establishes that the first disjunct of (42) is also false in w. $\square$

Proof of statement (46). Define total recursive f_e just as in the proof of (39), except that ${\mathsf{K}_i}^k\phi$ is replaced with ${\mathsf{K}_G}^k\phi$. For $j \in G_-$, define total recursive f_j just like f_e, but with the condition $L_c(\sigma|t', \mathsf{B}_i\,\phi) = 1)$ in place of condition $L_c(\sigma|t', \phi) = 1)$. Apply (64) to each f_i to obtain respective, total recursive function h_i. Let $\mathbf{h} = (h_1, \ldots, h_N)$.

Suppose that $\phi \in \Delta_e$ and that $\Delta_i \cap K_G^\omega = \emptyset$, for each $i \in G$. By the definition of $\mathbf{h}$, we have that for all $\mathbf{c} \in C^N$:

$$\mathbf{c} \;\equiv_{\boldsymbol{\Delta}}\; \mathbf{h}(\mathbf{c}); \tag{107}$$

so $\mathbf{h}$ preserves $\boldsymbol{\Delta}$. By construction, h depends only on $\boldsymbol{\Delta}$. Furthermore, for all $i \in G$, $z \in W$, $t \in T$, and $k \in \mathbb{N}$:

$$L_{h(c)}(s_{i,z}|t, @_{t^*}\neg{\mathsf{K}_G}^k\,\phi) = 1 \;\;\Leftrightarrow\;\; L_{h(c)}(s_{i,z}|t, @_{t^*}{\mathsf{K}_G}^k\,\phi) = 0;. \tag{108}$$

Suppose that $u \in I_{i,w,t^*}$ satisfies:

$$u \;\in\; \|\mathsf{IS5}_G\|^{t^*}_{\mathfrak{M}_{t^*}}; \tag{109}$$

$$u \;\in\; \|\mathsf{T}_{G,e}\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}; \tag{110}$$

$$u \;\in\; \|\mathsf{S}_{G,e}\boldsymbol{\Delta}\|^{t^*}_{\mathfrak{M}_{t^*}}; \tag{111}$$

$$u \;\in\; \|\mathsf{K}_{G,e}\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}. \tag{112}$$

Abbreviate:

$$\begin{aligned} \mathbf{c} &= \mathbf{c}_{i,u,t^*}; \\ x &= u[\mathbf{h}(\mathbf{c})/i, t^*]. \end{aligned}$$

From (107), (111) and (112), obtain via proposition 3 that:

$$x \;\in\; \|\mathsf{K}_{G,e}\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}. \tag{113}$$

Note that for $j \in G_-$ and $z \in W$ we have by the definition of $\mathbf{h}$ that:

$$L_{h_e(c_e)}(s_{i,z}|t, @_{t^*}{\mathsf{K}_G}^k\,\phi) = 1 \;\;\Leftrightarrow\;\; (\forall t' : t^* \leq t' \leq t)\, L_{h_e(c_e)}(s_{i,z}|t', @_{t^*}\,\phi) = 1; \tag{114}$$

$$L_{h_j(c_j)}(s_{i,z}|t, @_{t^*}{\mathsf{K}_G}^k\,\phi) = 1 \;\;\Leftrightarrow\;\; (\forall t' : t^* \leq t' \leq t)\, L_{h_j(c_j)}(s_{i,z}|t', @_{t^*}\mathsf{B}_{G,e}\,\phi) = 1; \tag{115}$$

Let $y \in D_{G,x,t^*} \subseteq I_{G,x,t^*}$. So $y \in I_{G,u,t^*}$ by (44). Then by (110), we have for all $j \in G_-$ that $y \in \|\mathsf{G}\tilde{\mathsf{C}}_j\mathsf{B}_e\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}$. Hence, by (114-115), we have for all $i \in G$, $y \in D_{G,x,t^*}$, and $k \in \mathbb{N}$:

$$L_{h_i(c_i)}(s_{i,y}|t, @_{t^*}{\mathsf{K}_G}^k\,\phi) = 1 \;\;\Leftrightarrow\;\; (\forall t' : t^* \leq t' \leq t)\, L_{h_e(c_e)}(s_{i,y}|t', @_{t^*}\,\phi) = 1; \tag{116}$$

By (113), (108), and (116), we have that $x \in \|\mathsf{K}_{G,j}\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}$, for all $j \in G_-$, so again by (113) we have $x \in \|\mathsf{K}_G{}^1\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}$, and hence, that $x \in \|\phi\|^{t^*}_{\mathfrak{M}_{t^*}} = \|\mathsf{K}_G{}^0\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}$. Thus, we have the base case $\|K^1_G(\phi)\|^{t^*}_{\mathfrak{M}_{t^*}}$.

Next, assume for induction that $x \in \|K^{k+1}_G(\phi)\|^{t^*}_{\mathfrak{M}_{t^*}}$ and show that $x \in \|K^{k+2}_G(\phi)\|^{t^*}_{\mathfrak{M}_{t^*}}$. By the induction hypothesis, we have, for each $i \in G$ that:

$$x \in \|\mathsf{K}_{G,i}\mathsf{K}_G{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}; \tag{117}$$

and, therefore:

$$x \in \|\mathsf{GC}_i\mathsf{K}_G{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}; \tag{118}$$

$$y \in \|\mathsf{FG}\neg\tilde{\mathsf{E}}\mathsf{K}_G{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}},\ \text{for all } y \in D_{G,x,t^*}\ . \tag{119}$$

For $x \in \|K^{k+2}_G(\phi)\|^{t^*}_{\mathfrak{M}_{t^*}}$, it suffices to show, for each $i \in G$, that: $x \in \|\mathsf{K}_{G,i}\mathsf{K}_{G,i}\mathsf{K}_G{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}$. For that, it suffices, in turn, to show:

$$x \in \|\mathsf{GC}_i\mathsf{K}_{G,i}\mathsf{K}_G{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}; \tag{120}$$

$$x \in \|\mathsf{FG}\neg\tilde{\mathsf{E}}\mathsf{K}_{G,i}\mathsf{K}_G{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}},\ \text{for all } y \in D_{G,x,t^*}\ . \tag{121}$$

Requirement (120) expands to the requirements:

$$x \in \|\mathsf{K}_{G,i}\mathsf{K}_G{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}; \tag{122}$$

$$x \in \|\mathsf{G[B]}_i\mathsf{K}_{G,i}\mathsf{K}_G{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}; \tag{123}$$

$$x \in \|\mathsf{G}\langle\mathsf{B}\rangle_i\mathsf{K}_{G,i}\mathsf{K}_G{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}. \tag{124}$$

Requirement (122) is just (117). Hence, (118) yields:

$$x \in \|\mathsf{G[B]}_i\mathsf{K}_G{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}; \tag{125}$$

$$x \in \|\mathsf{G}\langle\mathsf{B}\rangle_i\mathsf{K}_G{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}. \tag{126}$$

Requirements (123-124) follow from (125-126) and properties (108) and (116) of **h**.

For reuirement (121), suppose that $y \in D_{G,x,t^*}$. It suffices to show that for all $y \in D_{G,x,t^*}$:

$$y \in \|\mathsf{GF[B]}_i\neg\mathsf{K}_{G,i}\mathsf{K}_G{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}} \;\Rightarrow\; y \notin \|\mathsf{K}_{G,i}\mathsf{K}_G{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}; \tag{127}$$

$$y \in \|\mathsf{GF[B]}_i\mathsf{K}_{G,i}\mathsf{K}_G{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}} \;\Rightarrow\; y \in \|\mathsf{K}_{G,i}\mathsf{K}_G{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}. \tag{128}$$

For requirement (127), suppose that $y \in \|\mathsf{GF[B]}_i\neg\mathsf{K}_{G,i}\mathsf{K}_G{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}$. Then by properties (116) and (108) of h, we have that $y \notin \|\mathsf{G[B]}_i\mathsf{K}_G{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}$. So $y \notin \|\mathsf{K}_{G,i}\mathsf{K}_G{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}$.

For requirement (128), suppose that:

$$y \in \|\mathsf{GF[B]}_i\mathsf{K}_{G,i}\mathsf{K}_G{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}} \tag{129}$$

For the consequent $y \in \|\mathsf{K}_{G,i}\mathsf{K}_G{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}$, it suffices, as usual, to show the requirements:

$$y \in \|\mathsf{GC}_i\mathsf{K}_G{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}; \tag{130}$$

$$z \in \|\mathsf{FG}\neg\tilde{\mathsf{E}}_i\mathsf{K}_G{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}},\ \text{for all } z \in D_{G,y,t^*}\ . \tag{131}$$

Requirement (131) is just (119), since $D_{G,y,t^*} = D_{G,u,t^*}$ by (109).[35] Requirement (130) expands to:

$$y \in \|\mathsf{K}_G{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}; \tag{132}$$

$$y \in \|\mathsf{G}[\mathsf{B}]_i\mathsf{K}_G{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}; \tag{133}$$

$$y \in \|\mathsf{G}\langle\mathsf{B}\rangle_i\mathsf{K}_G{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}. \tag{134}$$

For requirement (132), we have from (129) and property (116) of h that $y \in \|\mathsf{GF}[\mathsf{B}]_i\mathsf{K}_G{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}$. So $y \in \|\mathsf{K}_G{}^k\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}$, by (119). For requirement (133), note that (129), along with property (116) of $\mathbf{h}$ implies that $y \in \|\mathsf{G}[\mathsf{B}]_i\,\phi\|^{t^*}_{\mathfrak{M}_{t^*}}$, which again, in light of property (116) implies requirement (133). Requirement (134) is then immediate by property (108) of $\mathbf{h}$. □

[35]This is the proof's only appeal to the S5 property for information.