Generics as Useful Defeasible Rules of Inference

Justin Helms

May 13, 2026

1 Introduction

Suppose scientists discover a new species of animal, the lorch. Twenty percent of lorches carry a disease called burlosis that can infect humans and causes serious symptoms. Would we accept that lorches carry burlosis? The empirical evidence suggests we would, despite only a fraction of lorches carrying the disease, because the acceptability of a generic sentence 'Fs G' is sensitive to the practical stakes of being wrong about whether a particular F is G. Now suppose all and only those lorches that carry the disease are bright purple, while those that don't are gray. Would we still accept that lorches carry burlosis? This paper presents evidence that we would not, and explains why.

Bare plural generics (Carlson, 1977) are sentences of the form 'Fs G', such as "mosquitos carry malaria" and "birds can fly." They have posed long-standing challenges to theories of natural language meaning. In this paper, we demonstrate a novel empirical phenomenon: Generic acceptability judgements are sensitive to the predictability of exceptions in a speaker's environment. In particular, speakers are more likely to accept 'Fs G' when the exceptions, those Fs that aren't G, are themselves predictable from what else the speaker knows. In our controlled-learning experiment, when a visible marker allowed participants to predict which blue mushrooms are poisonous, acceptance of "blue mushrooms are poisonous" fell from .50 to .05 ($p = .003$) and acceptance of "blue mushrooms are nutritious" rose from

.75 to 1.00 ($p = .047$). To capture this, we build a formal theory of generics inspired by a longstanding informal claim that generics encode useful defeasible rules of inference. Our model has a cross-validated R^2 of 0.923.

If the generic 'Fs G' is taken to mean that the prevalence $P(G \mid F)$ is higher than some threshold θ, then no one threshold seems adequate for capturing speakers' acceptability judgements. Consider, for example, that most ticks don't carry Lyme disease, yet speakers tend to accept "ticks carry Lyme disease," while most books are paperback, and yet speakers tend to reject "books are paperback." Tessler and Goodman (2019), however, found that if the threshold θ is taken to be a variable that's pragmatically determined at a context in accordance with Rational Speech Act (RSA) theory, then the probability that speakers will accept the generic can be accurately predicted in many cases. This is a cognitivist theory: It connects speakers' estimates of prevalence with their acceptance of generics.

What remains to be seen is how well the RSA model can be extended to predict the speakers' acceptance of generics as a function of the objective structure of their environment. We look at an end-to-end extension of RSA in which speakers are modelled as extracting the conditional prevalences of their decision environment from experience, and those prevalences are fed into the RSA model of generics. (We consider, and rule out, a fully Bayesian variant.)

We compare that baseline to our model of generics as useful defeasible rules of inference. Our model draws on a key insight of the RSA model: The probability of accepting 'Fs G' is a function of the pragmatic utility of the generic. In RSA, utility is taken to be a function of how informative the generic is about the prevalence. However, in our model pragmatic utility is the expected value of adopting a theory of which the rule of inference $F \to G$ is a part. This is an essentially holistic theory: The value of one rule of inference depends on what other rules of inference a speaker accepts. It is holism that allows the model to capture sensitivity to the predictability of exceptions. The expected value of a set of defeasible rules of inference, taken together, is modelled as a function of three things: (1) Epistemic value (the resulting inferences tend to be accurate); (2) practical value (the resulting inferences

tend to lead to actions with higher payoffs); and (3) simplicity (the total number of inferential rules is minimized).

The extended RSA model is fundamentally unable to capture sensitivity to the predictability of exceptions. Our model is able to capture this, and this is reflected in higher R^2 and lower AIC when evaluated against the full dataset. However, the results for leave-one-condition-out (LOCO) cross validation leave the picture less clear overall: Both models have a comparable cross-validated R^2. This points to the need for further data collection. In Section 2, we lay out the theory in formal detail. We first give background on the Tessler and Goodman (2019) model. We then build up our alternative conception of the pragmatic utility of a generic: Starting with the logic of defeasible inferences, working through our definition of a decision environment, the value of a default theory given an environment, and finally how probabilities are assigned to generics based on the value of theories to which the corresponding default rule belongs. In Section 3, we motivate and explain our experimental design. In Section 4, we give the results.

2 Theoretical Framework

Tessler and Goodman (2019) offers not just a formal theory of generics, but a quantitative theory which makes precise predictions about the probability of a speaker accepting a generic. Their model showed a strong fit with the animal kind generics they tested. However, there are three empirical facts which are unexplained by that model: First, generic acceptability judgements are sensitive to practical facts. 'Fs G' is sensitive to the importance of the property G, even when the prevalence of property G amongst Fs is established by the immediate context (Cimpian et al., 2010; Bian and Cimpian, 2021; Mirabile et al., 2024). Second, as the experiments from our paper show, generics are sensitive to the predictability of exceptions. Finally, Khemlani et al. (2012) demonstrated that generic acceptance, more so than prevalence estimates, predicted speakers' tendency to infer from something being a

particular kind to its having a particular property.

The RSA framework conceptualizes pragmatic utility as a function of informativeness. An utterance is taken to be true or false at a possible world, and is valuable to the extent that it leads the listener to assign higher probability to being in the actual world. This, in and of itself, is blind to practical stakes, but Sumers et al. (2024) offers an extension of the framework meant to capture their effect on speech. They model the listener as updating their policy π (a probability distribution over possible physical, non-speech acts) based on an utterance. The expected value of the listener's policy is added to an utterance's expected informativeness to determine its overall informativeness. They successfully modelled loose talk this way: When the stakes are low, a speaker might utter "it's 3pm" even when they know it's actually 3:01pm. However, when the stakes are higher, e.g. someone is setting their watch, the speaker might utter "it's 3:01pm" instead. In modelling generics, an RSA theorist might capture stakes-sensitivity in the same way.

However, it's harder to see how Tessler and Goodman (2019) can accommodate sensitivity to the predictability of exceptions. This sensitivity says that my acceptance of 'Fs G' is determined, in part, by whether or not I'll be able to tell which Fs aren't G. That suggests that the function of 'Fs G' isn't to inform me about the relationship between being an F and being a G, but to help me determine which Fs are G. Together with Khemlani et al. (2012), this suggests a deep connection between accepting a generic and making inferences.

That a generic encodes useful defeasible rules of inference is not a new idea, and has been explored in Stovall (2023); Restall (ming).[1] However, precisifying the claim in a way that's conducive to scientific inquiry has proven elusive. Going from that claim to exact predictions about the probability of a speaker accepting a generic requires three ingredients: A formal logic of defeasible inferences, a model of the value of a set of defeasible inferences, and a linking function telling us how likely it is a defeasible inference will be accepted given

[1]See also Pelletier and Asher (1997); Asher and Morreau (1990); Veltman (1996) for accounts of generics that both preserve a close connection between generics and defeasible rules of inference while still remaining largely compatible with possible world semantics

its relative value.

We model the value of an inferential rule as deriving from two sources: It has epistemic value insofar as it tends to make correct predictions; and it has practical value insofar as it tends to lead to actions with high material payoff. Then, given the value of a default rule, we model the probability that an agent will accept it using the Boltzmann linking function: Higher value rules have higher probability of being accepted in a manner that allows for a tradeoff between exploiting the agent's theory of the world and exploring alternatives.

But when applying the linking function, the holism raises an issue: The utility of one rule depends on the probability of the other rules. For example, if $F \to G$ is a rule with exceptions, and those exceptions can be captured (in whole or in part) by the rule $F \wedge H \to \neg G$, then the utility of $F \to G$ will be higher the higher the probability of $F \wedge H \to \neg G$. To cope with this dependency, we'll use the technique of mean field approximation to find a solution.

2.1 From the Value of an Action to the Probability of an Action

A foundational tenet of RSA is that speakers assign a pragmatic utility to each possible utterance, and then make that utterance in accordance with the Boltzmann linking function. That function is a model of how agents probabilistically choose an action a amongst possible actions A, given a utility function $V : A \to \mathbb{R}$:

$$P(a) = \frac{e^{\alpha \cdot V(a)}}{\sum_{a' \in A} e^{\alpha \cdot V(a')}} \tag{1}$$

One might think it would be better always to choose whichever action a has the highest pragmatic utility, $\bar{a} = \arg\max_{a \in A} V(a)$. However, there may be practical limits in our ability to do so, either because of the cognitive cost of computing $\arg\max_{a \in A} V(a)$, or because, even if computed, neural noise inhibits our ability to consistently act in line with that computation. Furthermore, there may be practical reasons not to always act in line with $\bar{a}$. $V(a)$ is the

subjective expected utility the agent assigns to an utterance. By choosing an utterance other than $\bar{a}$, an agent will be able to learn a better estimation of the expected utility of that alternative, which may, in fact, prove to be higher than $V(\bar{a})$. This is the exploitation-exploration tradeoff that agents face.

Exploration motivates agents to possibly choose an action a other than $\bar{a}$. We can ask what the expected value of the agent's action is, $\mathbb{E}_P[V(a)]$. If, for a given value of $\mathbb{E}_P[V(a)]$, an agent maximizes entropy, then P is given by the Boltzmann distribution (eq. 1). In that equation, α is the so-called inverse temperature. A lower α corresponds to more exploration, a lower probability of choosing $\bar{a}$, and a lower expected value of the action, $\mathbb{E}_P[V(a)]$. Once a particular value of $\mathbb{E}_P[V(a)]$ is fixed, choosing P so as to maximize entropy is a natural extension of the principle of indifference: To maximize entropy is to minimize the difference between P and the uniform distribution.[2]

The function has been found to predict the probability of human agents acting in a wide variety of experiments, serving as a link between subjective expected value and actual human decision-making. For example, Daw et al. (2006) shows that it predicts human agents' choices in multi-armed bandit environments.[3]

RSA has applied this model to speech acts. There, the relevant actions are utterances, and the relevant notion of value is informativeness. In our account, we preserve the use of the Boltzmann function for going from the value of an utterance to its probability, but the value of a generic derives from the inferences it licenses, rather than its informativeness.

[2]More particularly, maximizing the entropy minimizes the Kullback-Leibler divergence of the distribution P from the uniform distribution P_0. The Kullback-Leibler divergence is:

$$D_{KL}(P \parallel P_0) = \sum_{a \in A} P(a) \log \left(\frac{P(a)}{P_0(a)} \right)$$

[3]See also Icard (2023) and Lieder and Griffiths (2020) for a theoretical defense of its role in bounded rationality, and Sutton and Barto (2018) for a general discussion of its use in reinforcement learning.

2.2 An Overview of RSA

In RSA, typically, the communicative situation is idealized so that the speaker intends to communicate to a listener that they are in a particular world state w. The value of an utterance u is the log of the probability the listener will assign to being in w given that the utterance u is made:

$$V(u \mid w) = \log P_{L0}(w \mid u) \tag{2}$$

Here, P_{L0} is the subjective probability of a so-called literal listener. A literal listener, upon hearing u, puts probability mass on w if and only if $w \in [\![u]\!]$. So:

$$P_{L0}(w \mid u) \propto \mathbb{I}_{w \in [\![u]\!]} \cdot P_{prior}(w) \tag{3}$$

where $\mathbb{I}_{w \in [\![u]\!]}$ equals 1 if $w \in [\![u]\!]$ and 0 otherwise. $P_{prior}(w)$ is the probability the listener assigns to being in world-state w prior to updating on u. A so-called level-one pragmatic speaker will then utter u with probability $P_{S1}(u \mid w)$ given by plugging Eq. 2 into Eq. 1 (the Boltzmann linking function). From there, we can recurse: A level-one listener will assign probability to w not as a direct function of $[\![u]\!]$, but as a function of P_{S1}:

$$P_{L1}(w \mid u) \propto P_{S1}(u \mid w) \cdot P_{prior}(w) \tag{4}$$

This recursion is how RSA functions as a formalization of Gricean pragmatics. It allows us to model a speaker reasoning about a listener reasoning about the speaker, and so on. It has been found, for example, that different listeners are well modelled at different levels of depth in that recursion, some interpreting utterances in accordance with P_{L0}, some in accordance with P_{L1}, etc (Franke and Degen, 2016).

Tessler and Goodman (2019) offers a theory of generics within the RSA framework. A speaker is modelled as intending to communicate that the probability of a member of a kind

F having a property G is x, $P(G \mid F) = x$. A generic 'Fs G' is true just in case $P(G \mid F)$ is greater than some pragmatically determined threshold θ. In possible world semantics, we can write this as:

$$[\![\text{Fs G}]\!] = \{w : P(G \mid F) \geq \theta \text{ at } w\}$$

The speaker and listener have a shared prior P_{prior} over possible values of x. The prior is determined by a comparison class. For example, if the generic is "birds fly," the comparison class might be other animals, and P_{prior} is the probability that an animal flies.[4]

The speaker is choosing whether to accept the generic 'Fs G' or not. If the speaker accepts the generic, u_{gen}, then the literal listener will move all of their probability mass on to values of x that are greater than θ:

$$L_0(x \mid u_{gen}, \theta) = \begin{cases} \frac{P_{prior}(x)}{\int_\theta^1 P_{prior}(x')dx'} & \text{if } x > \theta \\ 0 & \text{otherwise} \end{cases}$$

If the speaker doesn't accept the generic, u_{null}, then the literal listener's probability won't change from the prior:

$$L_0(x \mid u_{null}, \theta) = P_{prior}(x)$$

However, the literal listener L_0 is underdetermined: Whether and how informative the generic will be will depend on the value θ. So the pragmatic listener L_1 is a model of the listener that resolves this ambiguity by marginalizing over θ:

[4]Specifically, the prior is modelled as a mixture of betas:

$$P_{prior}(x) = w_{low} \cdot \text{Beta}(x; 1, 100) + w_{high} \cdot \text{Beta}(x; 100, 1) + w_{mid} \cdot \text{Beta}(x; 1, 1)$$

Where w_{low}, w_{mid}, and w_{high} are the mixture weights summing to one. This will give us effectively two fitted parameters (since they sum to one we can fix one of them to be a function of the other two, $w_{mid} = 1 - w_{low} - w_{high}$).

$$L_1(x \mid u) = \int_0^1 L_0(x \mid u, \theta) P(\theta) d\theta$$

This is enough for the speaker to assign an informativeness value to each candidate utterance:

$$V(u \mid x) = \ln L_1(x \mid u)$$

Following standard RSA practice, we further charge an utterance cost $C(u)$, with $C(u_{null}) = 0$ and $C(u_{gen}) = c$, where c is a fitted parameter. The speaker's overall utility for an utterance is its informativeness minus the cost:

$$U(u \mid x) = V(u \mid x) - C(u)$$

These utilities are fed into the Boltzmann linking function (Eq. 1) to determine the probability that a speaker will accept the generic:

$$P_{accept}(\text{“Fs G”}) = \frac{\exp(\alpha \cdot U(u_{gen} \mid x))}{\sum_{u' \in \{u_{gen}, u_{null}\}} \exp(\alpha \cdot U(u' \mid x))}$$

where α is the speaker rationality parameter.

Tessler and Goodman (2019) is a cognitive model of generics: It predicts speakers' acceptance of generics as a function of their prevalence judgements. It does not predict generic acceptability judgements given the objective facts of the speakers' environment. As such, it is not, in and of itself, an alternative to the present model. Instead, as our baseline, we propose an end-to-end extension of Tessler and Goodman (2019): We model speakers as having extracted the conditional prevalence $x = P(G \mid F)$ from their decision environment, and we feed that value into the speaker model just described. The mixture-of-betas prior P_{prior} is retained, with its mixture weights fitted from data.[5]

[5]That is to say: For our baseline, we model the RSA speaker as having successfully extracted the objective conditional probabilities of the environment (the frequentist prevalence). Alternatively, we might have

2.3 Default Rules

On our theory, the value of a generic is in the epistemic value of the mental actions it leads to and the practical value of the physical actions it leads to. The relevant mental actions here are inferences.

To model defeasible rules of inference, we adapt the non-monotonic logic of Horty (2012). We use Horty's framework, which explicitly models prioritized defeat, rather than an unprioritized default logic like that of Reiter (1980). As we demonstrate in Section 4, Horty's concept of defeat is what allows our model to explain how speakers successfully shield general rules from the practical penalties of their specific exceptions.

Let $\mathcal{L}$ be a language of propositional logic generated by a finite set of atomic propositions $\mathbb{P}$. By ordinary propositions and ordinary beliefs, we'll mean well-formed formulas in $\mathcal{L}$. A default rule δ is a formula of the form $\phi \rightarrow \psi$ where ϕ and ψ are ordinary propositions. We write Premise(δ) and Conclusion(δ) to denote the left and right hand side of a default rule. For a set of default rules D, Premise(D) and Conclusion(D) are the set of premises of D and conclusions of D, respectively. In Horty (2012), a default theory is a triple, $\langle \mathcal{W}, D, < \rangle$, where $\mathcal{W}$ is a set of ordinary propositions, D is a set of default rules, and $<$ is a strict partial order of the default rules; the priority order tells us, if two defaults have inconsistent conclusions, which one to accept. However, we'll define a default theory as just a pair $\langle \mathcal{W}, D \rangle$, and assume that the strict partial order $<$ is induced.

Definition 1 (Default Theory)**.** A default theory is a tuple $\langle \mathcal{W}, D \rangle$ where:

1. $\mathcal{W}$ is a set of ordinary propositions.

2. D is a set of default rules.

Definition 2 (Induced Order)**.** The induced partial order $<$ of a default theory $\langle \mathcal{W}, D \rangle$ is

modelled the participant as a Bayesian learner updating a background prior. But we found that coupling a standard RSA mixture-prior with Bayesian updating results in extreme prior washout, causing the model to collapse. Thus, the frequentist-RSA model serves as the strongest possible prevalence-based competitor.

given by the following condition:

$$\phi_1 \to \psi_1 < \phi_2 \to \psi_2 \text{ if and only if } \mathcal{W} \cup \phi_2 \vdash \phi_1 \text{ and } \mathcal{W} \cup \phi_1 \nvdash \phi_2$$

This says that one rule has higher priority than another when its premise is strictly more specific. For example, taking $\delta_1 = bird \to fly$ and $\delta_2 = penguin \to \neg fly$ with $\mathcal{W}$ entailing $penguin \supset bird$, we have $\mathcal{W} \cup \{penguin\} \vdash bird$ but not the converse, so $\delta_1 < \delta_2$.

In a default theory, the critical task is to decide which conclusions of the default rules to accept, given the priority ordering and ordinary propositions $\mathcal{W}$. This is done by identifying stable scenarios: Subsets of D that are consistent with each other and $\mathcal{W}$. Spelling this out requires the following definitions:

Definition 3 (Triggered Defaults)**.** Given the default theory $\langle \mathcal{W}, D \rangle$ and a scenario $S \subseteq D$, the **triggered** defaults are:

$$\text{Triggered}(S) = \{\delta \in D \mid \text{Conclusion}(S) \cup \mathcal{W} \vdash \text{Premise}(\delta)\}$$

Definition 4 (Conflicted Defaults)**.** Given the default theory $\langle \mathcal{W}, D \rangle$ and a scenario $S \subseteq D$, the **conflicted** defaults are:

$$\text{Conflicted}(S) = \{\delta \in D \mid \text{Conclusion}(S) \cup \mathcal{W} \vdash \neg\text{Conclusion}(\delta)\}$$

Definition 5 (Defeated Defaults)**.** Given the default theory $\langle \mathcal{W}, D \rangle$ with induced order $<$ and a scenario $S \subseteq D$, the **defeated** defaults are:

$$\text{Defeated}(S) = \{\delta \in D \mid \exists \delta' \in \text{Triggered}(S) \text{ s.t. } \delta < \delta' \text{ and } \text{Conclusion}(\delta') \vdash \neg\text{Conclusion}(\delta)\}$$

A stable scenario S is then a kind of fixed point of a default theory: A subset of its defaults such that its members are all those defaults which are triggered by that subset

other than those which are conflicted or defeated:

Definition 6 (Stable Scenario)**.** $S \subseteq D$ of a default theory $\langle \mathcal{W}, D \rangle$ is a **stable scenario** if and only if:

$$S = \text{Triggered}(S) \setminus \text{Conflicted}(S) \setminus \text{Defeated}(S)$$

Let's work through a quick example. Let δ_1 be the default rule $bird \rightarrow fly$, which we can read as saying that something's being a bird defeasibly implies that it can fly. Let δ_2 be the default rule $penguin \rightarrow \neg fly$: Something's being a penguin defeasibly implies that it can't fly. Let $\mathcal{W} = \{penguin \supset bird, \textit{Tweety is a penguin}\}$, where $\supset$ is the material conditional; that is, an implication that can't be defeated. Now, consider the default theory $\langle \mathcal{W}, \{\delta_1, \delta_2\} \rangle$. Since $penguin \supset bird$, the induced order is just $\delta_1 < \delta_2$. There are four scenarios to consider: $\{\,\}$, $\{\delta_1\}$, $\{\delta_2\}$, and $\{\delta_1, \delta_2\}$. The scenario $\{\,\}$ isn't stable because it triggers δ_2, which is neither conflicted nor defeated. The scenarios $\{\delta_1\}$ and $\{\delta_1, \delta_2\}$ aren't stable because they include δ_1, which is defeated by δ_2, since δ_2 is higher priority and entails the negation of the conclusion of δ_1. That leaves $\{\delta_2\}$ as the only stable scenario.

Finally, the extension of a scenario S is the set of entailments of the conclusions of the rules in S along with propositions in $\mathcal{W}$: $\text{Ext}(S) = \{\phi \mid \text{Conclusion}(S) \cup \mathcal{W} \vdash \phi\}$. In our model, the value of a default theory derives from the content of the extensions of its stable scenarios. We propose that a generic 'Fs G' expresses the default rule $F \rightarrow G$. A speaker's willingness to accept the generic depends on the expected value of including that rule in their theory. To compute that value, we first need to connect the logic just described to the structure of the agent's environment.

2.4 Decision Environments

Let Ω be the set of all possible valuations for $\mathbb{P}$, the atomic propositions of our language of propositional logic, $\mathcal{L}$.

Definition 7 (Decision Environment)**.** A **decision environment** is a tuple $\mathcal{E} = \langle \Omega, P, \mathcal{O}, A, u \rangle$

where:

1. Ω is the set of all possible valuations for $\mathbb{P}$. We call $w \in \Omega$ a state.

2. P is a probability distribution over Ω.

3. $\mathcal{O}$ is a function from states to observations, $\mathcal{O} : \Omega \to \mathcal{P}(\mathcal{L})$.

4. A is a set of available actions.

5. u is a material payoff function, $u : \Omega \times A \to \mathbb{R}$.

We model the agent as choosing a background theory, $\langle \mathcal{W}, T \rangle$. We'll refer to T as the agent's background default theory. For simplicity, we'll assume $\mathcal{W}$ is fixed, and that the agent is choosing the rules in T from among those rules for which there is non-zero objective probability of the premise and conclusion both being true, which we'll call the plausible default rules and denote $\mathcal{D}$:

Definition 8 (Plausible Default Rules)**.** A default rule δ is **plausible** given a decision environment $\langle \Omega, P, \mathcal{O}, A, u \rangle$ if and only if:

$$P(\{w \mid w \vDash \text{Premise}(\delta) \wedge \text{Conclusion}(\delta)\}) > 0$$

For example, in the decision environments of our experiment, participants are shown mushrooms that are either red, yellow, or blue but never more than one. Therefore, the plausible defaults $\mathcal{D}$ do not include e.g. $yellow \to red$. The observations at each state w of a decision environment, together with the background theory, determines a state-specific theory, $\langle \mathcal{W} \cup \mathcal{O}(w), T \rangle$.

2.5 The Value of a Default Theory

The value of a default rule is evaluated in terms of the expected value of having it in the background default theory T, conditional on the premise of the rule being triggered. The

value of a background theory T flows from its stable scenarios. The value of the stable scenarios is derived from what atomic propositions they allow an agent to infer, and whether those inferences are correct or incorrect. How much value is derived from a particular inference is determined both by the material payoff of the environment, and by the agent's cognitive payoff. Agents are modelled as valuing correct inferences, disvaluing incorrect inferences, and disvaluing failing to infer whether to accept either an atomic proposition or its negation.[6] The extent to which they value/disvalue each of these things is determined by fitted parameters of the model, whereas the practical payoffs are facts of the environment.

Let V be the function that assigns a value to a theory conditional on a particular premise ϕ being triggered. Let Ω_ϕ be the set of worlds where ϕ is satisfied.

$$\mathrm{V}(T \mid \phi) = \sum_{w \in \Omega_\phi} P(w \mid \phi) \left[\mathrm{Epistemic}(T, w) + \mathrm{Practical}(T, w)\right] - \lambda_{comp} \cdot |T| \tag{5}$$

λ_{comp} is a fitted parameter reflecting a complexity penalty; all else being equal, agents prefer simpler theories. Note that minimizing complexity has been shown to shape natural language with respect to a wide variety of phenomena. Within the Rational Speech Act (RSA) framework, for example, speakers are modeled as utility-maximizing agents who trade off the informativity of an utterance against its production cost, frequently operationalized as word length or syntactic complexity (Frank and Goodman, 2012; Goodman and Frank, 2016). And in lexical semantics, the lexicon is shaped by a pressure to minimize cognitive complexity while maximizing informativeness, a tradeoff that successfully predicts the cross-linguistic structure of kinship categories and color terms (Kemp and Regier, 2012; Zaslavsky et al., 2018). More generally, there is a tendency toward simplicity in human categorization behavior, formalized in minimum description length accounts. In our model, the parameter λ_{comp} exerts a downward pressure on the size of the speaker's default theory.

[6]For a discussion of the philosophical project of assigning value to full beliefs based on accuracy, see Pettigrew (2016). For a philosophical defense of grounding rationality in measurements of accuracy, see Easwaran and Fitelson (2015). For an empirically grounded discussion of the human tendency to disvalue suspending judgement, see Kruglanski and Webster (1996).

First, let's break down the definition of Epistemic(T, w). The epistemic value of a theory is a function of not just how well it predicts what isn't at first observed, but also how well it would have predicted what is observed, had it not been observed. Therefore, to test the prediction of a literal p or $\neg p$, we have to temporarily remove p from the agent's observations, if present. Let $\mathcal{O}^{-p}(w)$ be the observation set of state w with any direct observation of p removed:

$$\mathcal{O}^{-p}(w) = \{\phi \in \mathcal{O}(w) \mid \phi \neq p\}$$

Let $\Gamma^{-p}(S, w)$ be the ordinary belief set generated by the stable scenario S based on these reduced observations:

$$\Gamma^{-p}(S, w) = \mathcal{O}^{-p}(w) \cup \text{Conclusion}(S)$$

We want to take the average value of the stable scenarios an agent will have at w given that p needs to be inferred. We denote the set of such scenarios $\mathbb{S}$:

$$\mathbb{S}^{-p}(T, w) = \text{ stable scenarios of } \langle \mathcal{W} \cup \mathcal{O}^{-p}(w), T \rangle$$

$$\text{Epistemic}(T, w) = \sum_{p \in \mathbb{P}} \frac{1}{|\mathbb{S}^{-p}(T, w)|} \sum_{S \in \mathbb{S}^{-p}(T, w)} \text{Score}(S, w, p)$$

And the score of S at w for p is then defined piecewise:

$$\text{Score}(S, w, p) = \begin{cases} \lambda_{\text{acc}} & \text{if } \Gamma^{-p}(S, w) \vdash p \text{ and } w \models p \quad \text{(True Positive)} \\ \lambda_{\text{acc}} & \text{if } \Gamma^{-p}(S, w) \vdash \neg p \text{ and } w \models \neg p \quad \text{(True Negative)} \\ -\lambda_{\text{err}} & \text{if } \Gamma^{-p}(S, w) \vdash p \text{ and } w \models \neg p \quad \text{(False Positive)} \\ -\lambda_{\text{err}} & \text{if } \Gamma^{-p}(S, w) \vdash \neg p \text{ and } w \models p \quad \text{(False Negative)} \\ -\lambda_{\text{sus}} & \text{if } \Gamma^{-p}(S, w) \nvdash p \text{ and } \Gamma^{-p}(S, w) \nvdash \neg p \quad \text{(Suspension)} \end{cases}$$

The parameters λ_{acc}, λ_{err}, and λ_{sus} are fitted: λ_{acc} is the reward for a correct prediction, λ_{err} is the penalty for an incorrect prediction, and λ_{sus} is the penalty for suspending judgement.

Next, we'll give the definition of Practical(T, w). The practical value of a theory at a state is the expected value of the agent's policy for acting given their theory. Let $\Gamma(S, w) = \mathcal{O}(w) \cup$ Conclusion(S) be the agent's propositional belief set given the scenario S and observations $\mathcal{O}(w)$ at w. Let $\sigma : \mathcal{P}(\mathcal{L}) \to \Delta(A)$ determine the agent's policy: A mapping from the agent's propositional beliefs to a probability distribution over actions. Then we can define the expected practical value of T at w as:

$$\text{Practical}(T, w) = \frac{1}{|\mathbb{S}(T, w)|} \sum_{S \in \mathbb{S}(T,w)} \sum_{a \in A} \sigma(a \mid \Gamma(S, w)) \cdot u(w, a)$$

The stable scenario S determines the ordinary belief set at a world state w, and that set is denoted $\Gamma(S, w)$. That belief set determines a probability distribution over actions, $\sigma(\cdot \mid \Gamma(S, w))$. The material payoff $u(w, a)$ is then summed, weighted by the probability of a given the policy σ. In our experiment, an agent can either *eat* or *discard* the mushroom at each trial. As a simplification, we assume that the agent will eat a mushroom if they infer that it's nutritious, discard it if it's poisonous, and do either with 50% probability if neither or both are inferred.

$$\sigma(a \mid \Gamma) = \begin{cases} 1 & \text{if } a = \text{discard and } \Gamma \vdash \mathit{Poisonous} \\ 1 & \text{if } a = \text{eat and } \Gamma \vdash \mathit{Nutritious} \\ 0.5 & \text{if } \Gamma \nvdash \mathit{Poisonous} \text{ and } \Gamma \nvdash \mathit{Nutritious} \\ 0.5 & \text{if } \Gamma \vdash \mathit{Poisonous} \text{ and } \Gamma \vdash \mathit{Nutritious} \end{cases}$$

This gives us everything we need to compute $\mathrm{V}(T \mid \phi)$, our measure of the expected value of a theory given that a particular premise is triggered.

To make this concrete, consider this counterexample to prevalence-based theories (from Leslie (2008)): "Dogs have three legs." Because three-legged non-domesticated animals tend not to survive in the wild, the probability that an animal has three legs given that it is a dog is elevated relative to its base rate across animals. Tessler and Goodman (2019) would therefore predict a relatively high probability of speakers accepting the generic, yet speakers tend to reject it. Our model naturally handles these problem cases through the interaction of expected value and defeasible logic. For "dogs have three legs," the practical cost of a false positive or false negative is low and relatively symmetric, but the overwhelming prevalence of four-legged dogs means the rule *dog* $\rightarrow$ *four-legged* is more useful. Should a speaker accept both *dog* $\rightarrow$ *four-legged* and *dog* $\rightarrow$ *three-legged*, the two rules would conflict whenever the premise *dog* is triggered. The speaker would therefore never get the epistemic benefit of the rule *dog* $\rightarrow$ *four-legged*. And if the speaker accepted *dog* $\rightarrow$ *three-legged* and rejected *dog* $\rightarrow$ *four-legged*, they'd get the epistemic value of that rule in those cases when a dog has three legs, but lose the epistemic value of *dog* $\rightarrow$ *four-legged*, which is higher in aggregate simply because four legged dogs are more prevalent.

2.6 The Mutual Dependence Problem

In our case, we are trying to predict the probability that an agent will accept a generic. However, in our model a generic is a default rule, and the value of a rule depends on what

other default rules the agent accepts. Therefore, we cannot determine the value of accepting a generic 'Fs G' without knowing the probability that the other generics are accepted. But we cannot know the probability that the other generics are accepted without knowing the probability that 'Fs G' is accepted. This has the shape of a mean field approximation problem (Blei et al., 2017), and we tackle it by treating it as such. We start by assuming that the probability of accepting each default rule is independent and that the agent is indifferent between accepting and rejecting the rule. Then we iteratively re-compute the value of accepting a rule given the current estimation of the probabilities of the other rules, then re-compute the probability of that rule as a function of the weighted sum of that value and the previously assigned probability.

Let $\pi^{(t)}(\delta)$ be the probability assigned to accepting default rule δ at iteration t. We assume the agent starts indifferent, so $\pi^{(0)}(\delta) = 0.5$. Then the probability of accepting a theory T at iteration t is:

$$P_{\pi^{(t)}}(T) = \prod_{\delta \in T} \pi^{(t)}(\delta) \prod_{\delta \notin T} (1 - \pi^{(t)}(\delta))$$

Then given a premise ϕ, at iteration t, the score of a theory is:

$$\mathcal{S}^{(t)}(T \mid \phi) = \alpha \cdot V(T \mid \phi) + w_{\text{prior}} \cdot \ln P_{\pi^{(t)}}(T)$$

Here, w_{prior} is a fitted parameter, the weight given to the current approximate probability distribution. $S^{(t)}$ is used to compute the updated probabilities of the theories, $P^{(t+1)}(T \mid \phi)$:

$$P^{(t+1)}(T \mid \phi) = \frac{\exp(\mathcal{S}^{(t)}(T \mid \phi))}{\sum_{T' \subseteq \mathcal{D}} \exp(\mathcal{S}^{(t)}(T' \mid \phi))}$$

From which we can recover the updated probabilities of the individual rules:

$$\pi^{(t+1)}(\delta) = \sum_{T : \delta \in T} P^{(t+1)}(T \mid \text{Premise}(\delta))$$

We iterate until convergence.

2.7 Core Predictions

The theory laid out above makes four core qualitative predictions about how a speaker's acceptance of a generic varies with the structure of their decision environment. We state them here, and will refer back to them by label going forward.

(P1) The probability of accepting a generic rises when the cost of a false negative rises. For a default rule $F \rightarrow G$, the expected value of including the rule in the agent's background theory rises with the practical cost of false negatives for that inference: the cost of failing to recognize an F as a G. So when G is a property whose recognition would prompt an action that averts a cost, acceptance of 'Fs G' should be higher in environments where that cost is greater, all else equal. In these cases, the $Practical(T, w)$ term in Eq. 5 will be lower for all theories T that don't contain $F \rightarrow G$ at worlds w where both F and G are true. This prediction is in line with the striking-property hypothesis of Leslie (2008) and related theoretical (van Rooij and Schulz, 2020) and empirical work (Cimpian et al., 2010), which predicts that generic acceptability judgements are sensitive to the strikingness of the property attributed to a kind.

(P2) The probability of accepting a generic falls when the cost of a false positive rises. Conversely, the expected value of $F \rightarrow G$ falls with the practical cost of false positives for that inference: the cost of acting as though an F is G when it isn't. So when G is a property whose attribution would license an action that is costly in those cases where the attribution is incorrect, acceptance of 'Fs G' should be lower in environments where that cost is greater, all else equal. In these cases, the $Practical(T, w)$ term in Eq. 5 will be lower for all theories T that do contain $F \rightarrow G$ at worlds w where F is true and G is false. This prediction does not follow from the striking-property hypothesis: the relevant cost is the

cost of false positives, not the cost of failing to detect a striking feature, and the property G itself need not become more or less striking across conditions.[7]

(P3) The probability of accepting a generic rises when its exceptions are predictable. Suppose the exceptions to $F \to G$ (the Fs that aren't G) are themselves predictable from a more specific feature H, so that the rule $F \wedge H \to \neg G$ is available. Because the premise of $F \wedge H \to \neg G$ is strictly more specific than that of $F \to G$, the induced order gives $F \to G < F \wedge H \to \neg G$. Where $F \to G$ would have led to a false inference, it is now defeated by $F \wedge H \to \neg G$, which provides a correct inference. Therefore, in these cases, the $Epistemic(T, w)$ and $Practical(T, w)$ terms in Eq. 5 will be higher at worlds w where F, H, and $\neg G$ are true, for all T that contain both $F \to G$ and $F \wedge H \to \neg G$. This raises the overall expected value of $F \to G$, and so acceptance of 'Fs G' should be higher in environments in which such an H is observable.

(P4) The probability of accepting a generic falls when its true positives are otherwise predictable. Conversely, suppose the Fs that are G are themselves predictable from a more specific feature H, so that the rule $F \wedge H \to G$ is available and more statistically reliable (leads to fewer false inferences). Where $F \to G$ would have led to a true inference, it is now superfluous: the higher-priority $F \wedge H \to G$ would have drawn the same conclusion. $F \to G$ therefore makes a distinctive contribution to the agent's inferences only in those cases when it would have led to a false inference. Therefore, in these cases, the $Epistemic(T, w)$ and $Practical(T, w)$ terms in Eq. 5 will be no higher at worlds w where F, H, and G are true, for T that contain both $F \to G$ and $F \wedge H \to G$ compared to theories that contain only $F \wedge H \to G$. At the same time, the theory that contains both will have lower relative value because of the complexity penalty, $-\lambda_{comp} \cdot |T|$. This lowers $F \to G$'s overall expected

[7] A defender of strikingness-based theories might respond by re-defining strikingness as itself sensitive to the cost of errors. That is, a property is more striking if false negatives are costly, and less striking if false positives are costly. This would, however, dissolve much of the explanatory content of the original proposal, since strikingness would then be defined in terms of a decision-theoretic framework, which would ultimately be driving the predictions.

value, and so acceptance of 'Fs G' should be lower in environments in which such an H is observable.

Predictions (P1) and (P2) concern the sensitivity of generic acceptability to practical stakes; (P3) and (P4) concern its sensitivity to the predictability of exceptions. None of (P1)–(P4) can be captured by an end-to-end extension of the Tessler and Goodman (2019) RSA model that takes only the objective conditional prevalence $P(G \mid F)$ and prior as input, since each of the four predictions concerns a comparison in which $P(G \mid F)$ and the prior can be held fixed.

3 Experimental Design: The Two-Armed Bandit

As formalized by the mutual dependence problem above, our theory is inherently holistic: The expected value of accepting any single generic generalization depends heavily on the agent's background default theory and the other generics they currently accept. This holism presents a significant methodological challenge. Traditional empirical research on generics often involves collecting acceptability judgments for real-world kinds. However, evaluating our theory against real-world kinds is highly vulnerable to confounders.

For example, the acceptability of a generic "Fs G" has been hypothesized to be sensitive to the strikingness of the property G and to the cue validity of the generic, $P(F \mid G)$. Many generics that speakers accept even though they have low prevalence both have high cue validity and involve striking properties. For instance, consider the generic "mosquitos carry malaria." Speakers accept this generic, and it's both true that carrying malaria is a striking property and that cue validity is high (the probability that something is a mosquito given that it carries malaria is high). Which of these two facts drives our acceptance of the generic? Or do both? A controlled learning environment is needed. In the present theory, the acceptability of one generic can be higher or lower in virtue of another generic being higher or lower. If this holism is correct, then for real-world kinds, it may be difficult

to know which generic acceptability judgements might interact with each other and survey speakers in a way that exhausts all possibilities.

To isolate the mechanics of the defeasible logic, we elicited generic acceptability judgements for artificial kinds whose statistical structure and material payoffs participants had learned in a preceding foraging task. In the first phase, participants learned about the properties of artificial mushrooms through a two-armed bandit task in which eating a nutritious mushroom was rewarded and eating a poisonous one was penalized. In the second phase, participants indicated whether they agreed or disagreed with generic statements about those mushrooms. This design departs from the dominant paradigm in empirical work on generic acceptability (Cimpian et al., 2010; Brandone et al., 2012; Tessler and Goodman, 2019), which conveys prevalence information to participants through vignettes or stipulated percentages. Allowing the statistical structure of the environment to be acquired through experience, rather than asserted, makes it possible to manipulate material stakes and the predictability of exceptions while holding objective prevalence fixed, and avoids presupposing any particular causal relationship between prevalence estimates and generic acceptance; two judgements that Cimpian et al. (2010) showed to dissociate, with generics accepted at prevalences well below those that generics are taken to imply. Methodologically, the foraging phase adapts standard paradigms from probabilistic category learning (Knowlton et al., 1994; Ashby and Maddox, 2005) and multi-armed bandit tasks (Daw et al., 2006).

In the foraging task, participants are shown an image of a mushroom. The mushrooms can be red, blue, or yellow, and in the marked conditions, blue mushrooms additionally vary in whether they are spotted (Figure 1). Prevalence rates of poisonous/nutritious mushrooms by color are held constant across conditions: 90% of red mushrooms are poisonous, 0% of yellow mushrooms are poisonous, and 10% of blue mushrooms are poisonous. There are 20 trials with each color, for 60 total trials. At each trial, participants must choose whether to eat or discard the mushroom they're shown. Participants start with a $3.00 USD bonus. Whenever they eat a nutritious mushroom, their bonus increases by $0.05 USD. Across

conditions, the penalty for eating a poisonous mushroom varies. In high stakes conditions, eating a poisonous mushroom decreases the participant's bonus by \$1.00 USD; in medium stakes, the penalty is \$0.50 USD; in low stakes, it's \$0.05 USD.

(a) Blue (b) Spotted blue (c) Red (d) Yellow

Figure 1: The four mushroom stimuli used in the foraging task. In marked conditions, poisonous blue mushrooms are visually distinguished by spots.

The five experimental conditions are summarized in Table 1. Conditions vary along two dimensions: the penalty for eating a poisonous mushroom (\$0.05, \$0.50, or \$1.00) and whether poisonous blue mushrooms are visually marked with spots.

Table 1: Summary of experimental conditions. The 'generics' column gives the number of generic statements each participant evaluated.

Condition	Notation	Penalty	Visible Marking	Generics	Participants
Low stakes	$\mathcal{E}_l$	\$0.05	No	6	20
Medium stakes	$\mathcal{E}_m$	\$0.50	No	6	20
High stakes	$\mathcal{E}_h$	\$1.00	No	6	20
Marked high stakes	$\mathcal{E}_{mh}$	\$1.00	Yes	10	20
Marked low stakes	$\mathcal{E}_{ml}$	\$0.05	Yes	10	20

The atomic propositions will be denoted $\mathbb{P} = \{Red, Blue, Yellow, Spotted, Poisonous, Nutritious\}$, e.g. *Red* can be interpreted as meaning that the mushroom is red. We'll denote the states as w with subscripts for the true atomics. E.g., w_{rp} is the state where the mushroom is red and poisonous. Table 2 gives the probability of each state in each decision environment. Note that the probability of a mushroom being a certain color given that it's poisonous or given that it's nutritious is the same in all conditions. E.g., $P(Red \mid Poisonous)$ is the same in all conditions, as is $P(Poisonous \mid Red)$. This experiment is not meant to establish the role of

cue validity or prevalence in generic acceptability judgements. Rather, it holds these fixed and varies only the material payoff and the predictability of exceptions in the environment.

Table 2: Probability $P(w)$ of each state in each decision environment. Each color appears with equal frequency ($\frac{1}{3}$ of trials). Unmarked environments ($\mathcal{E}_l$, $\mathcal{E}_m$, $\mathcal{E}_h$) share the same probability distribution, as do marked environments ($\mathcal{E}_{mh}$, $\mathcal{E}_{ml}$), where poisonous blue mushrooms are spotted.

State	$\mathcal{E}_l, \mathcal{E}_m, \mathcal{E}_h$	$\mathcal{E}_{mh}, \mathcal{E}_{ml}$
w_{rp}	$\frac{3}{10}$	$\frac{3}{10}$
w_{rn}	$\frac{1}{30}$	$\frac{1}{30}$
w_{bp}	$\frac{1}{30}$	0
w_{bsp}	0	$\frac{1}{30}$
w_{bn}	$\frac{3}{10}$	$\frac{3}{10}$
w_{yn}	$\frac{1}{3}$	$\frac{1}{3}$

Instantiating prediction (P1) of Section 2.7, our theory predicts that the percentage of participants who accept "blue mushrooms are poisonous" will be higher in higher-stakes conditions, since the cost of false negatives for the inference *blue mushrooms* $\rightarrow$ *poisonous* is higher in those conditions. Instantiating prediction (P2) of Section 2.7, our theory predicts that the percentage who accept "blue mushrooms are nutritious" will be lower in higher-stakes conditions, since the cost of false positives for the inference *blue mushrooms* $\rightarrow$ *nutritious* is higher in those conditions. This second prediction is the diagnostic one against the striking-property hypothesis: the nature of being nutritious does not vary across conditions; eating a nutritious mushroom is always worth exactly \$0.05 USD to a participant.

Conditions also vary in the predictability of exceptions. In the marked conditions, those blue mushrooms which are poisonous are also spotted, so that the rule $\delta_s =$ *spotted blue mushrooms* $\rightarrow$ *poisonous* is available; in the unmarked conditions, no such rule is available. Instantiating prediction (P3) of Section 2.7, the probability of accepting "blue mushrooms are nutritious" should be higher in the marked conditions, since the exceptions to that generic are defeated in the marked conditions by δ_s. And by (P4), the probability of accepting "blue mushrooms are poisonous" should be lower in the marked conditions, since the cases in which that generic

is true are captured by δ_s, leaving the generic to make a distinctive contribution only in the cases in which it is false.

4 Results

We collected acceptability judgments from $N = 20$ participants for each of the five conditions (100 total participants), yielding 760 total observations. To isolate the specific mechanisms driving these judgments, we compare our full model (expected value derived from the information value of a prioritized default logic) against two robust baselines: The unprioritized baseline (which retains expected value but removes the defeat mechanism, functioning as a standard default logic) and an extended-RSA baseline (which uses ground-truth-prevalence paired with a pragmatic speaker model, evaluating prevalence but blind to material payoff and the predictability of exceptions).

Figure 2 plots the model's predicted acceptability rates against observed acceptability rates across all conditions and generic types. The model achieves an overall $R^2 = 0.964$ with a mean absolute error of 0.064 across 38 comparisons.

Table 3 reports the predicted and observed acceptability rates for each generic statement in each condition.

Table 3: Predicted and observed acceptability rates (Pred / Obs) for each generic statement by condition. Dashes indicate generics not tested in that condition.

	Pred / Obs				
Generic	Low	Medium	High	Mkd. low	Mkd. high
Yellow mushrooms are poisonous	.08 / .10	.08 / .00	.08 / .20	.10 / .00	.09 / .00
Yellow mushrooms are nutritious	.97 / .95	.97 / 1.00	.97 / .95	.96 / .90	.97 / 1.00
Red mushrooms are poisonous	.96 / 1.00	1.00 / .95	1.00 / .95	.93 / .95	1.00 / 1.00
Red mushrooms are nutritious	.12 / .10	.00 / .10	.00 / .10	.14 / .05	.00 / .00
Blue mushrooms are poisonous	.12 / .20	.24 / .45	.52 / .50	.16 / .20	.13 / .05
Blue mushrooms are nutritious	.96 / .90	.89 / .80	.67 / .75	.89 / .80	.92 / 1.00
Blue mushrooms are spotted	—	—	—	.23 / .11	.11 / .20
Spotted mushrooms are blue	—	—	—	.78 / .84	.78 / .95
Spotted mushrooms are poisonous	—	—	—	.94 / .84	1.00 / 1.00
Spotted mushrooms are nutritious	—	—	—	.08 / .05	.00 / .00

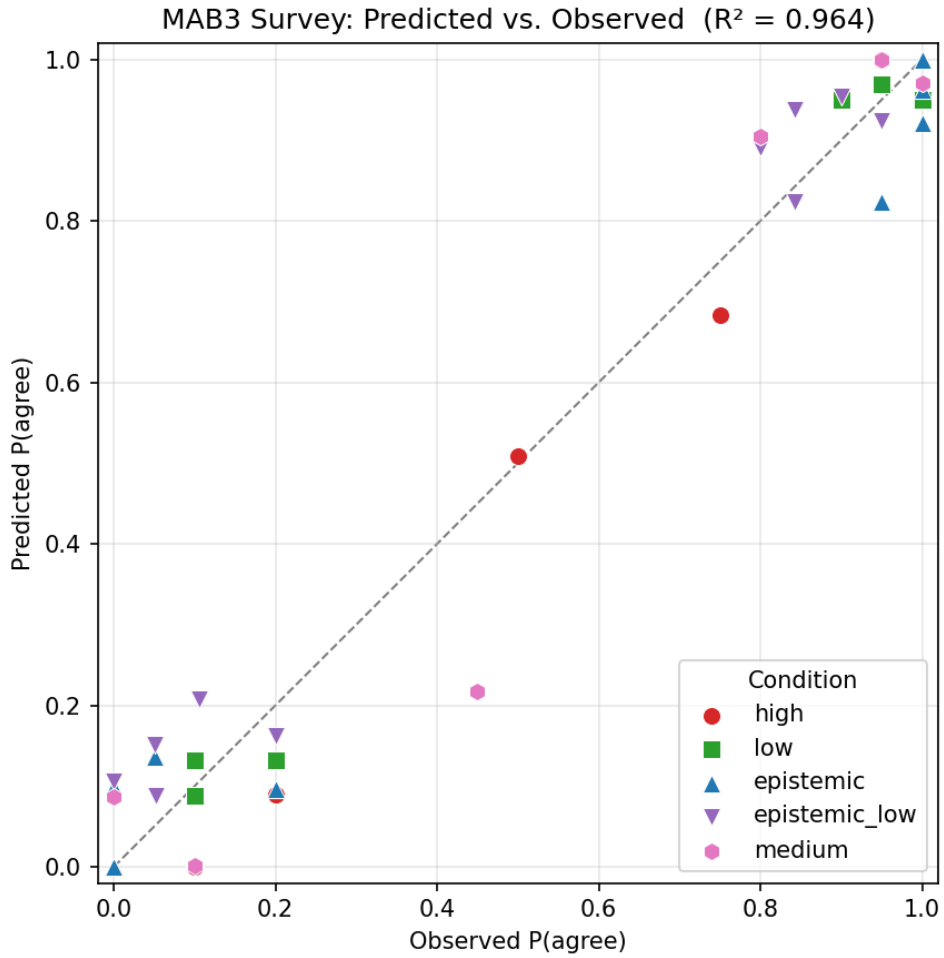

Figure 2: Predicted vs. observed acceptability rates for all generic types across all five conditions. Each point represents one generic–condition pair. The dashed line indicates perfect prediction.

Table 4 reports the fitted parameter values for the full model. The penalty for suspending judgment (λ_{sus}) is the largest epistemic parameter, followed by the penalty for inaccuracy (λ_{err}) and the reward for accuracy (λ_{acc}). This ordering indicates that agents are most averse to failing to draw an inference, followed by drawing an incorrect one, and least motivated by the reward for drawing a correct one.

To examine the key qualitative predictions directly, Figure 3 isolates the four critical pairwise comparisons for blue mushrooms. Panels (a) and (c) test predictions (P1) and (P2), respectively, with prevalence held constant; panels (b) and (d) test predictions (P4)

Table 4: Fitted parameter values for the full model (all five conditions).

Parameter	Description	Value
λ_{acc}	Reward for correct inference	3.047
λ_{err}	Penalty for incorrect inference	5.171
λ_{sus}	Penalty for suspending judgment	9.143
λ_{comp}	Complexity penalty	3.715
w_{prior}	Prior weight	0.759
α	Inverse temperature	0.037

and (P3), respectively, with stakes held constant.

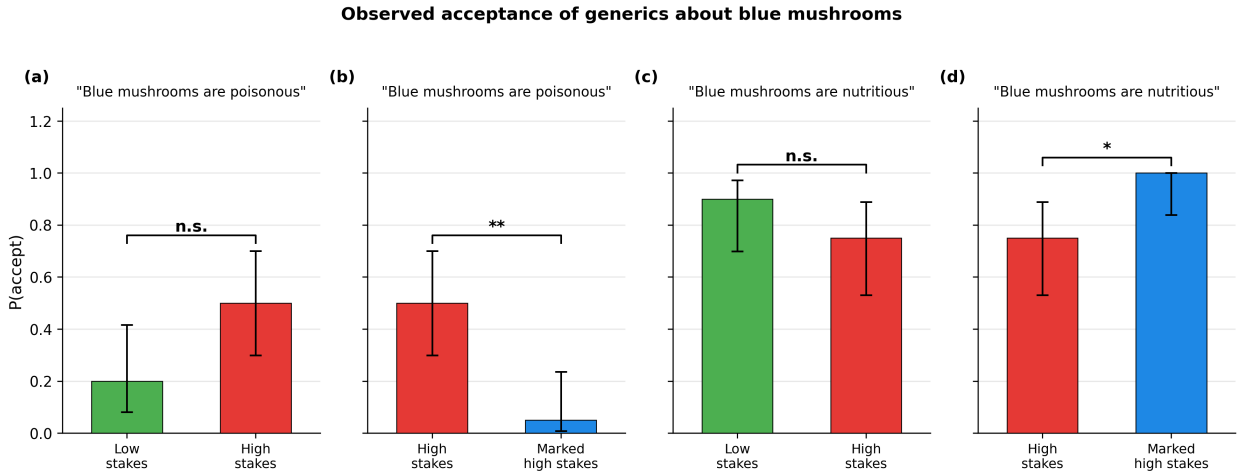

Figure 3: Observed acceptance rates for "blue mushrooms are poisonous" and "blue mushrooms are nutritious" across conditions. Error bars show 95% Wilson score confidence intervals. Significance brackets report two-sided Fisher's exact tests: * $p < .05$, ** $p < .01$, *** $p < .001$, n.s. $p \geq .05$.

The effect of exception predictability is clear: Introducing a visible marker for poisonous blue mushrooms dramatically reduces acceptance of "blue mushrooms are poisonous" (from .50 to .05, $p = .003$) and increases acceptance of "blue mushrooms are nutritious" (from .75 to 1.00, $p = .047$). Both effects are statistically significant and consistent with predictions (P3) and (P4).

The effects predicted by (P1) and (P2) are in the predicted direction: Acceptance of "blue mushrooms are poisonous" increases numerically from .20 to .50 as stakes rise, and acceptance of "blue mushrooms are nutritious" decreases from .90 to .75. But neither pairwise

comparison reaches conventional significance at $n = 20$ per group ($p = .096$ and $p = .408$, respectively). A post-hoc power analysis (assuming the observed proportions reflect the true effect sizes) indicates that the "poisonous" comparison has a medium-to-large effect size (Cohen's $h = 0.64$) for which $n = 38$ per group would yield 80% power, while the "nutritious" comparison has a smaller effect size (Cohen's $h = 0.40$) requiring approximately $n = 97$ per group. The current study, with $n = 20$ per group, had only 53% and 25% power to detect these effects, respectively. Thus, the absence of statistical significance for the stakes manipulation likely reflects insufficient power rather than the absence of an effect, particularly for the "poisonous" comparison where the observed 30-percentage-point difference is substantial.

To assess the generalizability of the model, we performed leave-one-condition-out cross-validation: For each of the five conditions, we fit the model on the remaining four conditions and evaluated predictions on the held-out condition. Table 5 reports the results. The model achieves an overall cross-validated $R^2 = 0.923$ (mean held-out $R^2 = 0.901$), indicating that the model generalizes well across conditions. Notably, however, it performs relatively poorly when holding out the medium stakes condition ($R^2 = 0.884$) and especially the high stakes condition ($R^2 = 0.747$). The model struggles, in particular, with predicting the correct value for "blue mushrooms are poisonous" when holding these conditions out. Note, too, that in the model fitted on all conditions (see Table 3), "blue mushrooms are poisonous" in the medium stakes condition has the largest discrepancy between the observed and predicted probability of acceptance (predicted is .24, observed is .45, for a difference of 0.21). There are two possibilities: First, with only 20 participants per condition, the 95% confidence interval for that observation is $[.23, .68]$. The poor fit on that data point may just fall out of the study being underpowered. Second, there is considerable evidence that the value of money to human agents is non-linear, whereas we model that value as linear.

Table 5: Leave-one-condition-out cross-validation results. For each fold, the model is trained on four conditions and evaluated on the held-out condition.

Held-out Condition	Held-out R^2	Held-out LL	Train R^2
High stakes	0.747	−63.5	0.960
Low stakes	0.986	−35.8	0.933
Marked high stakes	0.955	−32.3	0.930
Marked low stakes	0.933	−71.3	0.940
Medium stakes	0.884	−43.6	0.949
Overall CV	0.923	−246.5	—

4.1 Baseline: Extended RSA with Frequentist Prevalence

As described in Section 2.2, we compare against an end-to-end extension of the Tessler and Goodman (2019) RSA model of generics, using frequentist prevalence estimates drawn from the objective conditional probabilities of the decision environment. This model has four fitted parameters: the inverse temperature parameter α, the utterance cost c, and the mixture weights w_{low} and w_{high} (with $w_{mid} = 1 - w_{low} - w_{high}$). Figure 4 plots predicted against observed acceptability rates. The model achieves $R^2 = 0.934$ with a mean absolute error of 0.078.

Table 6: Fitted parameter values for the extended RSA frequentist baseline.

Parameter	Description	Value
α	Speaker rationality	1.260
c	Utterance cost	0.194
w_{low}	Low-prevalence mixture weight	0.832
w_{high}	High-prevalence mixture weight	0.027

The RSA baseline achieves a respectable overall fit, but its limitations are revealing. Because the model's predictions are driven entirely by prevalence, it assigns the same predicted acceptability to a given generic across all conditions that share the same prevalence structure. In particular, for the three unmarked conditions ($\mathcal{E}_l$, $\mathcal{E}_m$, $\mathcal{E}_h$), which have identical prevalence rates but differ in material payoff, the model predicts the same acceptance rate for "blue mushrooms are poisonous": 0.172 across all three. The observed rates, however,

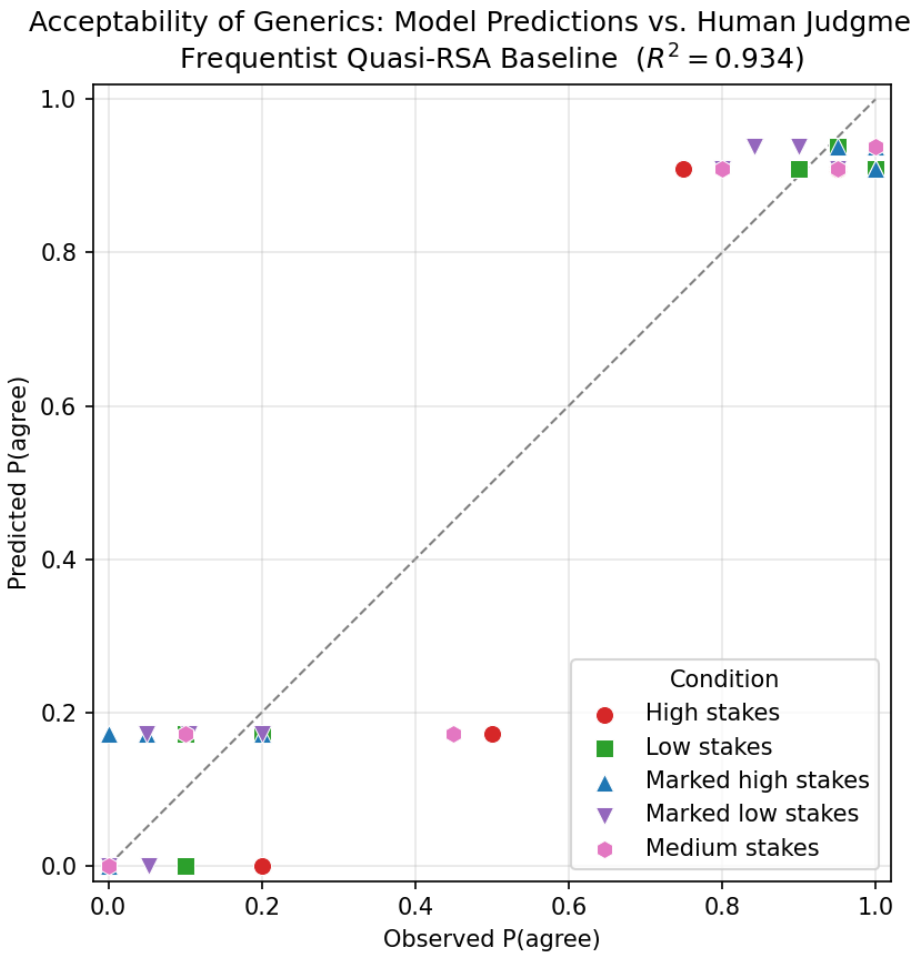

Figure 4: Predicted vs. observed acceptability rates for the extended RSA baseline with frequentist prevalence ($R^2 = 0.934$). Each point represents one generic–condition pair. The dashed line indicates perfect prediction.

are 0.20, 0.45, and 0.50 for low, medium, and high stakes, respectively. The RSA model is structurally unable to capture the effect of material stakes on generic acceptability. This is visible in Figure 4 as horizontal bands of points at a single predicted value spanning a wide range of observed values.

Leave-one-condition-out cross-validation yields an overall $R^2 = 0.920$ (mean held-out $R^2 = 0.893$). Table 7 reports the per-fold results. Performance is weakest when holding out the high stakes condition ($R^2 = 0.709$) and the medium stakes condition ($R^2 = 0.881$), mirroring the full model's difficulty with these conditions but to a greater degree.

Table 7: Leave-one-condition-out cross-validation for the extended RSA frequentist baseline.

Held-out Condition	Held-out R^2	Held-out LL	Train R^2
High stakes	0.709	−68.3	0.958
Low stakes	0.974	−39.0	0.926
Marked high stakes	0.954	−33.2	0.919
Marked low stakes	0.948	−72.6	0.926
Medium stakes	0.881	−42.7	0.942
Overall CV	0.920	−255.8	—

4.2 Ablation: Unprioritized Default Logic

To isolate the contribution of prioritized defeat, we compare the full model against an ablation that replaces Horty's prioritized default logic with an unprioritized default logic structurally identical to Reiter (1980). In the unprioritized variant, when two triggered defaults have conflicting conclusions, the resulting stable scenarios are those that include exactly one of the conflicting defaults; neither defeats the other, and the value function averages over both scenarios. This model retains the expected-value framework and has the same six fitted parameters as the full model. Figure 5 plots the results. The model achieves $R^2 = 0.927$ with a mean absolute error of 0.084.

Table 8: Fitted parameter values for the unprioritized default logic ablation.

Parameter	Description	Value
λ_{acc}	Reward for correct inference	7.381
λ_{err}	Penalty for incorrect inference	83.901
λ_{sus}	Penalty for suspending judgment	44.318
λ_{comp}	Complexity penalty	0.307
w_{prior}	Prior weight	0.500
α	Inverse temperature	0.013

The unprioritized model's failure is most apparent in the marked conditions, where it cannot exploit the predictability of exceptions. Without defeat, accepting both *blue* → *nutritious* and *spotted blue* → *poisonous* leads to suspension whenever a spotted blue mushroom is encountered: The two rules conflict, and since neither has priority, neither con-

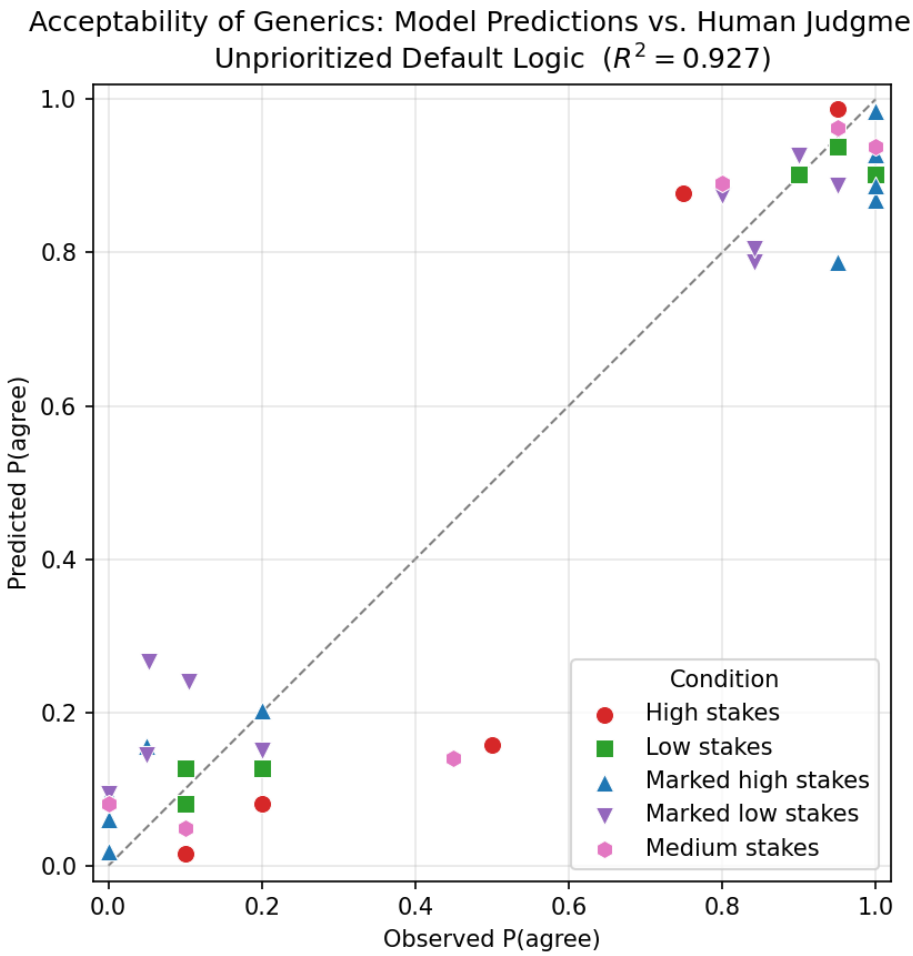

Figure 5: Predicted vs. observed acceptability rates for the unprioritized default logic ablation ($R^2 = 0.927$). Each point represents one generic–condition pair. The dashed line indicates perfect prediction.

clusion is drawn. By contrast, in the full model with prioritized defeat, the more specific rule *spotted blue* → *poisonous* defeats the general rule *blue* → *nutritious*, allowing the agent to correctly infer poisonous for spotted blue mushrooms while preserving the general rule's value for non-spotted blue mushrooms. This difference manifests most clearly in the marked low stakes condition: The unprioritized model predicts "spotted mushrooms are nutritious" at 0.266, far above the observed rate of 0.053, whereas the full model predicts 0.080. The unprioritized model is also forced into extreme parameter values to compensate: The inaccuracy penalty ($\lambda_{\text{err}} = 83.9$) is over sixteen times the accuracy reward ($\lambda_{\text{acc}} = 7.4$), which

suggests the optimizer is straining to approximate prioritized defeat through a high penalty asymmetry.

Leave-one-condition-out cross-validation yields an overall $R^2 = 0.900$ (mean held-out $R^2 = 0.880$). Table 9 reports the per-fold results.

Table 9: Leave-one-condition-out cross-validation for the unprioritized default logic ablation.

Held-out Condition	Held-out R^2	Held-out LL	Train R^2
High stakes	0.717	−65.8	0.951
Low stakes	0.982	−36.3	0.918
Marked high stakes	0.929	−38.3	0.909
Marked low stakes	0.896	−73.2	0.930
Medium stakes	0.874	−46.3	0.942
Overall CV	0.900	−259.8	—

4.3 Model Comparison

Table 10 summarizes the three models. The full model with prioritized defeat achieves the best fit on every metric despite having the same number of fitted parameters as the unprioritized ablation. The extended baseline, with only four parameters, achieves a competitive R^2 but is structurally unable to capture the effect of material stakes. The unprioritized ablation captures stakes sensitivity but cannot exploit the predictability of exceptions (failing to capture (P3) and (P4)), resulting in a worse fit than the full model and even worse cross-validation performance than the RSA baseline despite having more parameters.

Table 10: Summary comparison of the three models. Params is the number of fitted parameters. LL is the in-sample log-likelihood. AIC $= 2k - 2\text{LL}$. CV R^2 is the overall cross-validated R^2.

Model	Params	R^2	LL	AIC	CV R^2
Defeasible (full model)	6	0.964	−228.4	468.8	0.923
Extended RSA (frequentist)	4	0.934	−244.9	497.8	0.920
Unprioritized default logic	6	0.927	−243.0	498.0	0.900

5 General Discussion

There are two important takeaways from these results. First, they establish the empirical reality of sensitivity to the predictability of exceptions. To our knowledge, this phenomenon has not been explicitly identified in prior empirical work on generics, plausibly because it is only visible when one can manipulate whether the exceptions to a candidate generic are themselves predictable, which requires the kind of controlled artificial environment that we use here. Second, they show that a conception of generics as inferences can be made scientifically viable.

In the past, inference-based theories of generics have been formulated at the level of qualitative principles and illustrative examples, without the quantitative precision needed to test them against graded acceptability data. Non-monotonic logics provide a flexible apparatus that can, for any given generic, be configured to predict its acceptance or rejection after the fact; but flexible post-hoc accommodation is not prediction. Our theory and methodology are meant to bring the discussion beyond the exchange of intuitions about such accommodations. In assigning precise values to default rules relative to a given environment and background theory, and then assigning probabilities to such values, we seek to do for inference-based accounts of generics what Tessler and Goodman (2019) did for prevalence-based accounts. And by testing generics for kinds introduced through a learning environment, we can precisely control for objective prevalence and material payoffs.

The phenomenon of sensitivity to the predictability of exceptions is established by the data; and an inference-based account of generics has been given a form in which it can be quantitatively tested against graded acceptability judgements. What these results do not yet decisively establish is the superiority of our inference-based model over the extended prevalence-based one. The prevalence-based model and the inference-based model look comparable under leave-one-condition-out cross validation. A higher-powered version of these experiments needs to be run. And more conditions are needed. For example, in this experiment the overall prevalence of poisonous/nutritious mushrooms has been fixed across

conditions. So these conditions can't test how well the model captures sensitivity to cue validity. Furthermore, additional conditions with intermediary material payoffs and prevalences could help clarify any subtle ways these features of environment might drive generic acceptability.

Regardless, structurally, the end-to-end extension of the RSA model cannot capture either sensitivity to the predictability of exceptions or sensitivity to practical payoffs. However, each half of this model might be modified without being entirely abandoned. The frequentist learner was chosen for the baseline because it outperformed a Bayesian learner that jointly learned a mixture-of-betas distribution with the RSA model. But there are other options. In fact, we know that the high cost of false negatives systematically causes participants to overestimate the probability of an event (Harris et al., 2009). A model that captures this, such as Lieder et al. (2018), might be used instead of the frequentist learner. Capturing the sensitivity to the predictability of exceptions might prove less tractable, but here too it might be possible to distort the prior in a way that produces what's needed.

Furthermore, there are long standing conceptual reasons for rejecting the notion that a generic is an inferential rule. Namely, a default rule $F \rightarrow G$ cannot be assigned a fixed context-change potential, if we take the context-change potential to be a set of possible worlds. That is, if a generic just encoded a default rule $F \rightarrow G$, then how $F \rightarrow G$ changed the private information state of a listener would vary depending on the background theory of the listener. For a listener who doesn't know anything about the relationship between being a bird and being able to fly, learning birds $\rightarrow$ fly would update their private information state to reject worlds in which ostriches can't fly. Whereas, for a listener who already has the rule ostrich $\rightarrow \neg$fly, accepting birds $\rightarrow$ fly wouldn't lead to that same update in private information state.

This is an issue because it points to an embedding problem. If we can't assign fixed sets of possible worlds to generics, it becomes a challenge to say how the meaning of a generic contributes to the meaning of complex sentences in which it's embedded. If "penguins can

fly" encodes a default rule, then what is the meaning of, say, "either penguins can fly or their wings have another function" or "if penguins can fly, then Tweety will be able to escape." One path forward might be that of Cariani (ming). That paper offers a synthesis of possible world semantics and default theories of reasoning, whereby the set of possible worlds that a sentence encodes is constrained by a set of default rules.

A more radical path forward would be to build on the proof theoretic semantics of Francez (2015), on which the meaning of sentence is given by the inferences that license it and that it licenses, rather than by a set of possible worlds. While Francez's logic is monotonic, it might be extended to accommodate defeasible rules of inference.

6 Conclusion

Generic acceptability is sensitive to the predictability of exceptions. When a visible marker allowed participants in our learning experiment to determine which blue mushrooms were poisonous, acceptance of "blue mushrooms are poisonous" fell from .50 to .05, and acceptance of "blue mushrooms are nutritious" rose from .75 to 1.00. To our knowledge this phenomenon has not been documented in prior work on generics, plausibly because it becomes visible only when one can experimentally manipulate whether the exceptions to a candidate generic are themselves predictable.

Capturing this sensitivity requires moving from a prevalence-based account to an inference-based one. We have formalized the longstanding informal idea that generics encode useful defeasible rules of inference, and shown that such a model can make precise, graded predictions of acceptability, doing for inference-based accounts what Tessler and Goodman (2019) did for prevalence-based ones. In our data, the model achieves $R^2 = 0.964$ and outperforms an end-to-end extension of RSA that, structurally, cannot capture sensitivity to either material payoffs or the predictability of exceptions. This fit, however, isn't decisive: under leave-one-condition-out cross-validation the two models perform comparably (0.923 vs.

0.920). Adjudicating between accounts will require higher-powered experiments and a wider range of conditions, including conditions that vary cue validity and conditions with intermediate prevalences and payoffs. The present contribution is to put a previously undocumented empirical phenomenon on the table, and to put a previously informal theoretical tradition into a form where it can be quantitatively tested.

References

Ashby, F. G. and Maddox, W. T. (2005). Human category learning. *Annu. Rev. Psychol.*, 56(1):149–178.

Asher, N. and Morreau, M. (1990). Commonsense entailment: A modal theory of nonmonotonic reasoning. In *European Workshop on Logics in Artificial Intelligence*, pages 1–30. Springer.

Bian, L. and Cimpian, A. (2021). Generics about categories and generics about individuals: Same phenomenon or different? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 47(11):1836.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

Brandone, A. C., Cimpian, A., Leslie, S.-J., and Gelman, S. A. (2012). Do lions have manes? for children, generics are about kinds rather than quantities. *Child development*, 83(2):423–433.

Cariani, F. (forthcoming). Anankastic conditionals and the default theory of reasons. Manuscript.

Carlson, G. N. (1977). A unified analysis of the english bare plural. *Linguistics and philosophy*, 1(3):413–457.

Cimpian, A., Brandone, A. C., and Gelman, S. A. (2010). Generic statements require little evidence for acceptance but have powerful implications. *Cognitive Science*, 34(8):1452–1482.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., and Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, 441(7095):876–879.

Easwaran, K. and Fitelson, B. (2015). Accuracy, coherence, and evidence. *Oxford studies in epistemology*, 5:61–96.

Francez, N. (2015). *Proof-theoretic semantics*, volume 573. College Publications London.

Frank, M. C. and Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Franke, M. and Degen, J. (2016). Reasoning in reference games: Individual-vs. population-level probabilistic modeling. *PloS one*, 11(5):e0154854.

Goodman, N. D. and Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818–829.

Harris, A. J., Corner, A., and Hahn, U. (2009). Estimating the probability of negative events. *Cognition*, 110(1):51–64.

Horty, J. F. (2012). *Reasons as Defaults*. Oxford University Press.

Icard, T. F. (2023). Resource rationality.

Kemp, C. and Regier, T. (2012). Kinship categories across languages reflect general cognitive principles. *Science*, 336(6084):1049–1054.

Khemlani, S., Leslie, S.-J., and Glucksberg, S. (2012). Inferences about members of kinds: The generics hypothesis. *Language and Cognitive Processes*, 27(6):887–900.

Knowlton, B. J., Squire, L. R., and Gluck, M. A. (1994). Probabilistic classification learning in amnesia. *Learning & memory*, 1(2):106–120.

Kruglanski, A. W. and Webster, D. M. (1996). Motivated closing of the mind: "Seizing" and "freezing". *Psychological Review*, 103(2):263–283.

Leslie, S.-J. (2008). Generics: Cognition and acquisition. *Philosophical Review*, 117(1):1–47.

Lieder, F. and Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43:e1.

Lieder, F., Griffiths, T. L., and Hsu, M. (2018). Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological review*, 125(1):1.

Mirabile, P., Van Rooij, R., and Schulz, K. (2024). The role of impact on the meaning of generic sentences. *Frontiers in Psychology*, 15:1363390.

Pelletier, F. J. and Asher, N. (1997). Generics: An introduction. *Journal of Semantics*, 14(2):125–164.

Pettigrew, R. (2016). *Accuracy and the Laws of Credence*. Oxford University Press.

Reiter, R. (1980). A logic for default reasoning. *Artificial intelligence*, 13(1-2):81–132.

Restall, G. (forthcoming). Generics: Inference & accommodation. In Haslanger, S., Jones, K., Schroeter, F., and Schroeter, L., editors, *Mind, Language, and Social Hierarchy: Constructing a Shared Social World*. Oxford University Press.

Stovall, P. (2023). Characterizing generics are material inference tickets: A proof-theoretic analysis. *Inquiry*, 66(5):668–704.

Sumers, T. R., Ho, M. K., Griffiths, T. L., and Hawkins, R. D. (2024). Reconciling truthfulness and relevance as epistemic and decision-theoretic utility. *Psychological review*, 131(1):194.

Sutton, R. S. and Barto, A. G. (2018). Reinforcement learning: An introduction.

Tessler, M. H. and Goodman, N. D. (2019). The language of generalization. *Psychological Review*, 126(3):395–436.

van Rooij, R. and Schulz, K. (2020). Generics and typicality: A bounded rationality approach. *Linguistics and Philosophy*, 43(1):83–117.

Veltman, F. (1996). Defaults in update semantics. *Journal of philosophical logic*, 25(3):221–261.

Zaslavsky, N., Kemp, C., Regier, T., and Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, 115(31):7937–7942.