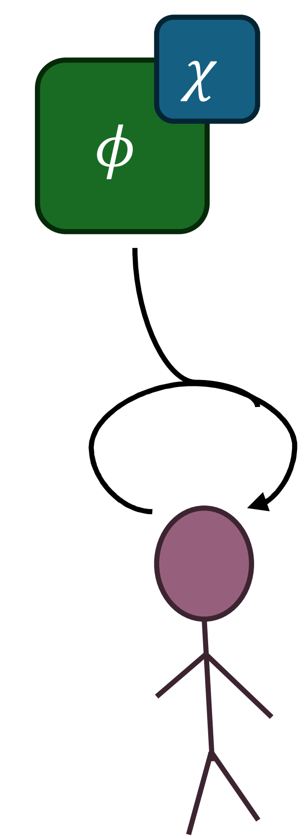
Learning with Confidence

Oliver Richardson

Uncertainty in Artificial Intelligence (UAI) 2025

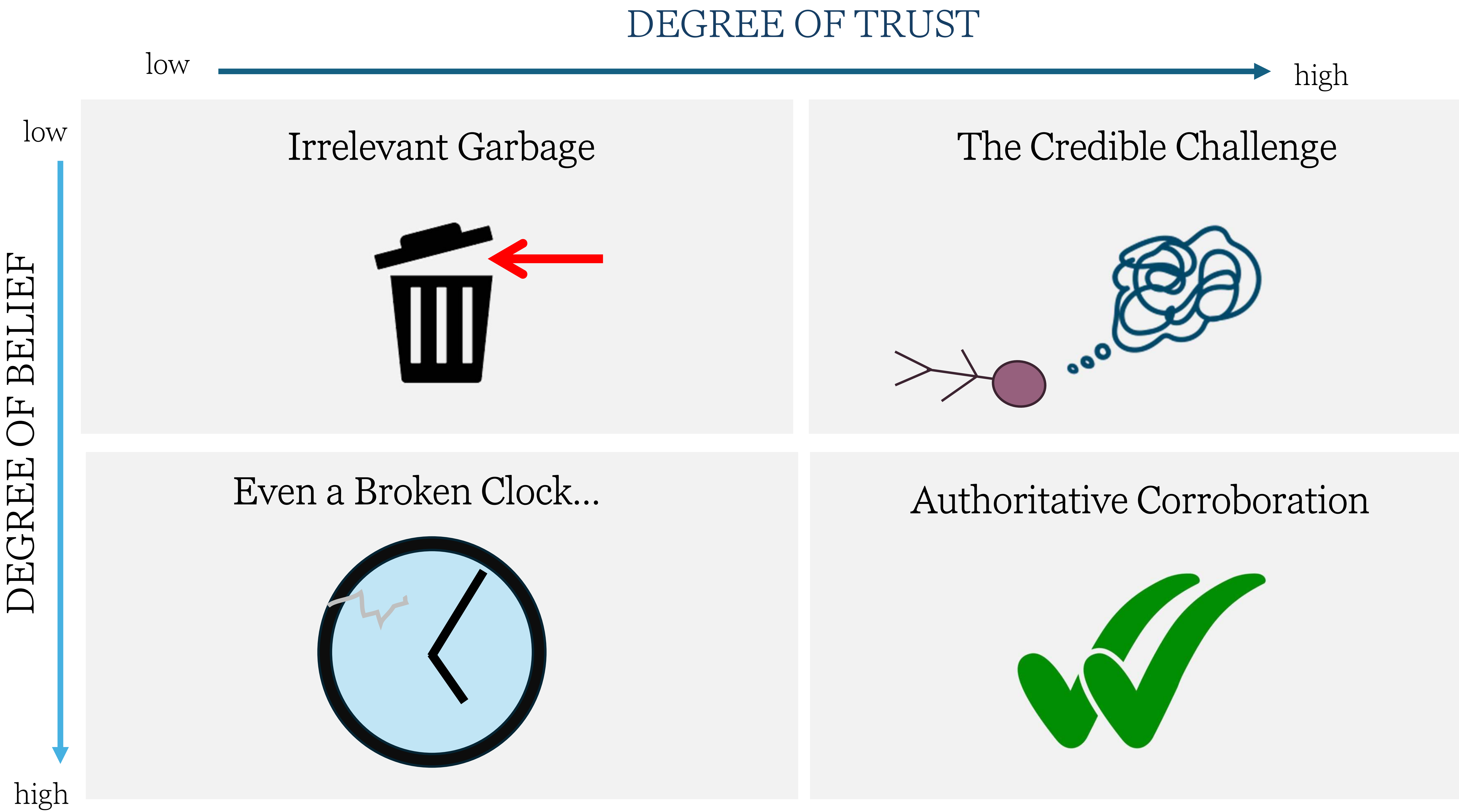
Mila / Université de Montréal

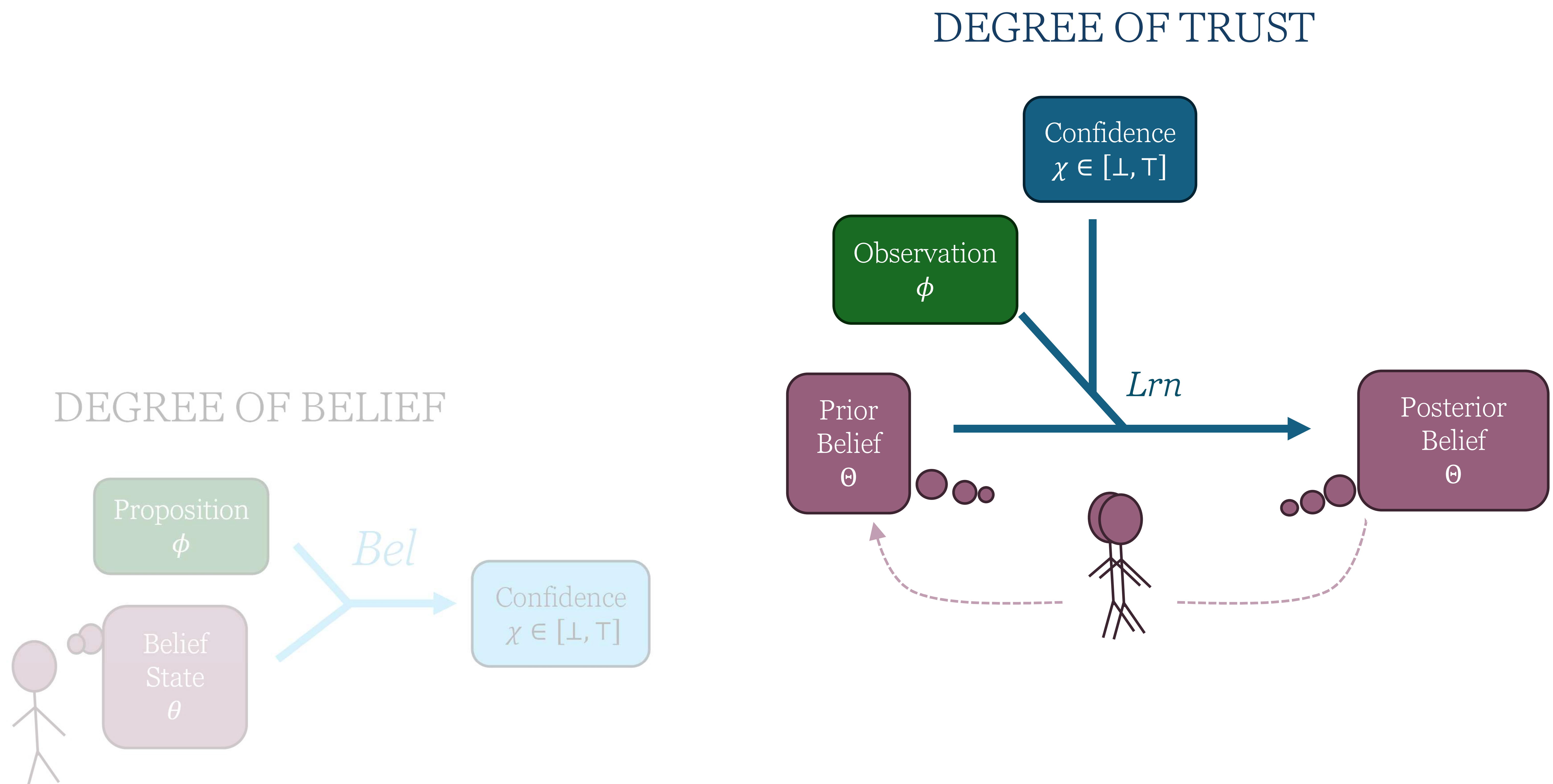


What does it mean (not) to have *confidence* in a statement ϕ?

Two interpretations:

- How likely do I find it? — DEGREE OF BELIEF
- How much should it influence my beliefs? — DEGREE OF TRUST


DEGREE OF TRUST
low
high
DEGREE OF BELIEF
low
high
Irrelevant Garbage
The Credible Challenge
Even a Broken Clock...
Authoritative Corroboration



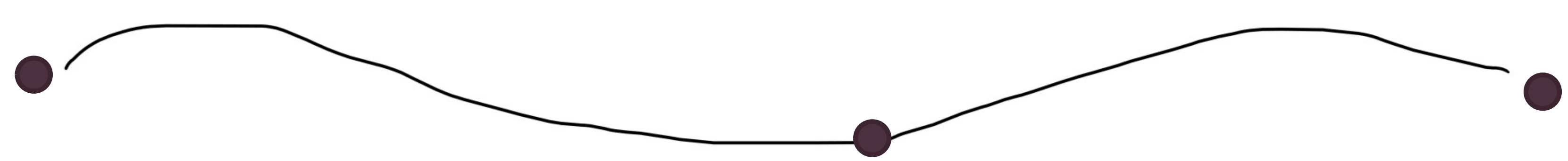
DEGREE OF TRUST
Confidence
$\chi \in [\perp, \top]$
Observation
ϕ
Lrn
Prior Belief
Θ
Posterior Belief
Θ
DEGREE OF BELIEF
Proposition
ϕ
Bel
Confidence
$\chi \in [\perp, \top]$
Belief State
θ

A Simple Example: Linear Interpolation

- belief states $\mu \in \Theta = \Delta(W)$ are probability measures;
- statements $A \in \Phi \subseteq W$ are events;
- confidence $\chi \in [0, 1]$ in the unit interval;

Notes:
- no obvious probabilistic interpretation of χ?
- full-confidence update is a projection

$$Lrn(A, \chi, \mu) = (1 - \chi)\mu + \chi(\mu|A)$$



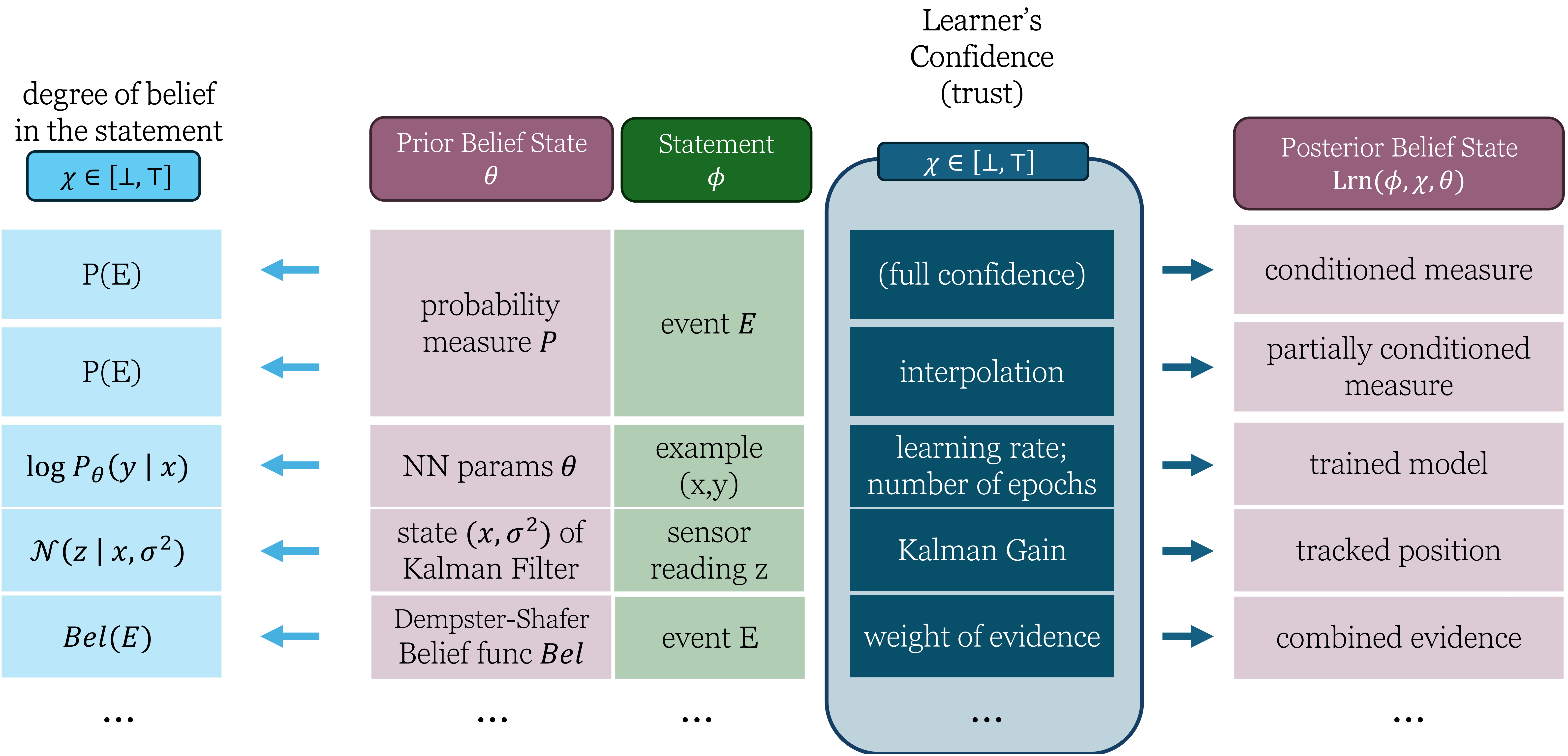
ignore @ no confidence

$$Lrn(A, \bot, \mu) = \mu$$

fully incorporate @ full confidence

$$Lrn(A, \top, \mu) = \mu|A$$

Unifying Existing Concepts



Confidence Domain

$$[\bot, \top] = (D, \leq, \oplus, \top, \bot, \mathfrak{g})$$

$\subset D \times D$

$: D \times D \to D$

$\in D$

preorder

no confidence

full confidence

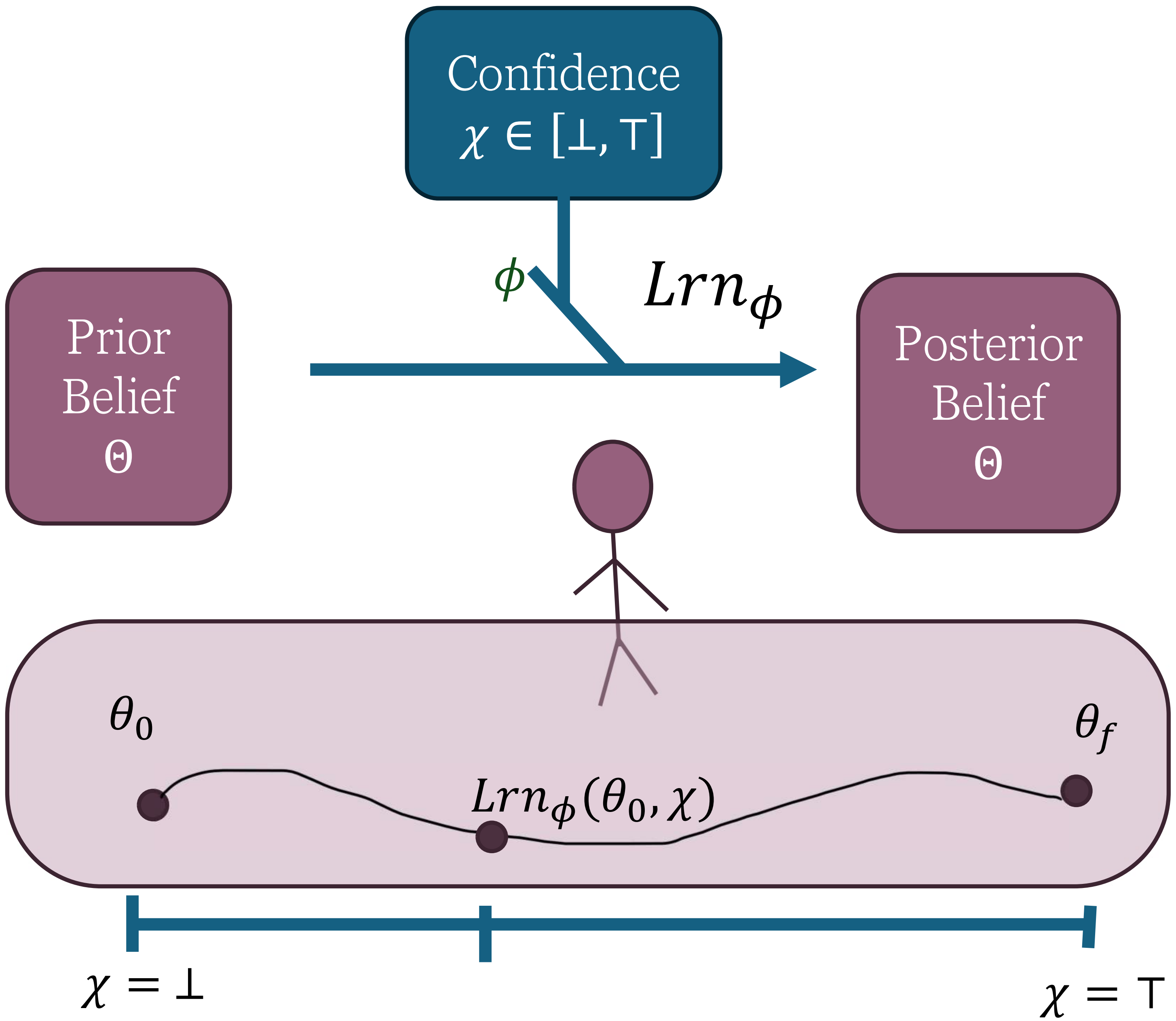
geometry

(topology, diffble structure on D)

independent combination

$$(\chi \oplus \chi') \oplus \chi'' = \chi \oplus (\chi' \oplus \chi'') \qquad \text{(associativity),}$$
$$\bot \oplus \chi = \chi \qquad \text{(that } \bot \text{ is neutral),}$$
$$\top \oplus \chi = \top \qquad \text{(and that } \top \text{ is absorbing).}$$

Axioms for Confidence



no confidence [L1] $Lrn_\phi(\bot, \theta) = \theta.$

full confidence [FC] $Lrn_\phi^\top \circ Lrn_\phi^\top = Lrn_\phi^\top.$

continuity [L2] $\chi \mapsto Lrn(\theta, \chi, \phi)$ is continuous, twice diffble

residuals [L3] $\chi < \chi' \implies \exists \chi''. \bot < \chi'' \leq \chi' \; Lrn_\phi^{\chi''} \circ Lrn_\phi^{\chi}(\theta) = Lrn_\phi^{\chi'}(\theta).$

acyclic [L4] If $\chi_0 \leq \chi \leq \chi_1$ and $Lrn_\phi(\chi_0, \theta) = Lrn_\phi(\chi_1, \theta)$, then $Lrn_\phi(\chi, \theta) = Lrn_\phi(\chi_0, \theta)$.

combinative [L5] $Lrn_\phi(\chi, Lrn_\phi(\chi', \theta)) = Lrn_\phi(\chi \oplus \chi', \theta)$

An *action* of the confidence domain

$$\left(D, \bot, \top, \mathfrak{g}, \leq, \oplus, \smile \right)$$

no confidence [L1] $Lrn_\phi(\bot, \theta) = \theta.$

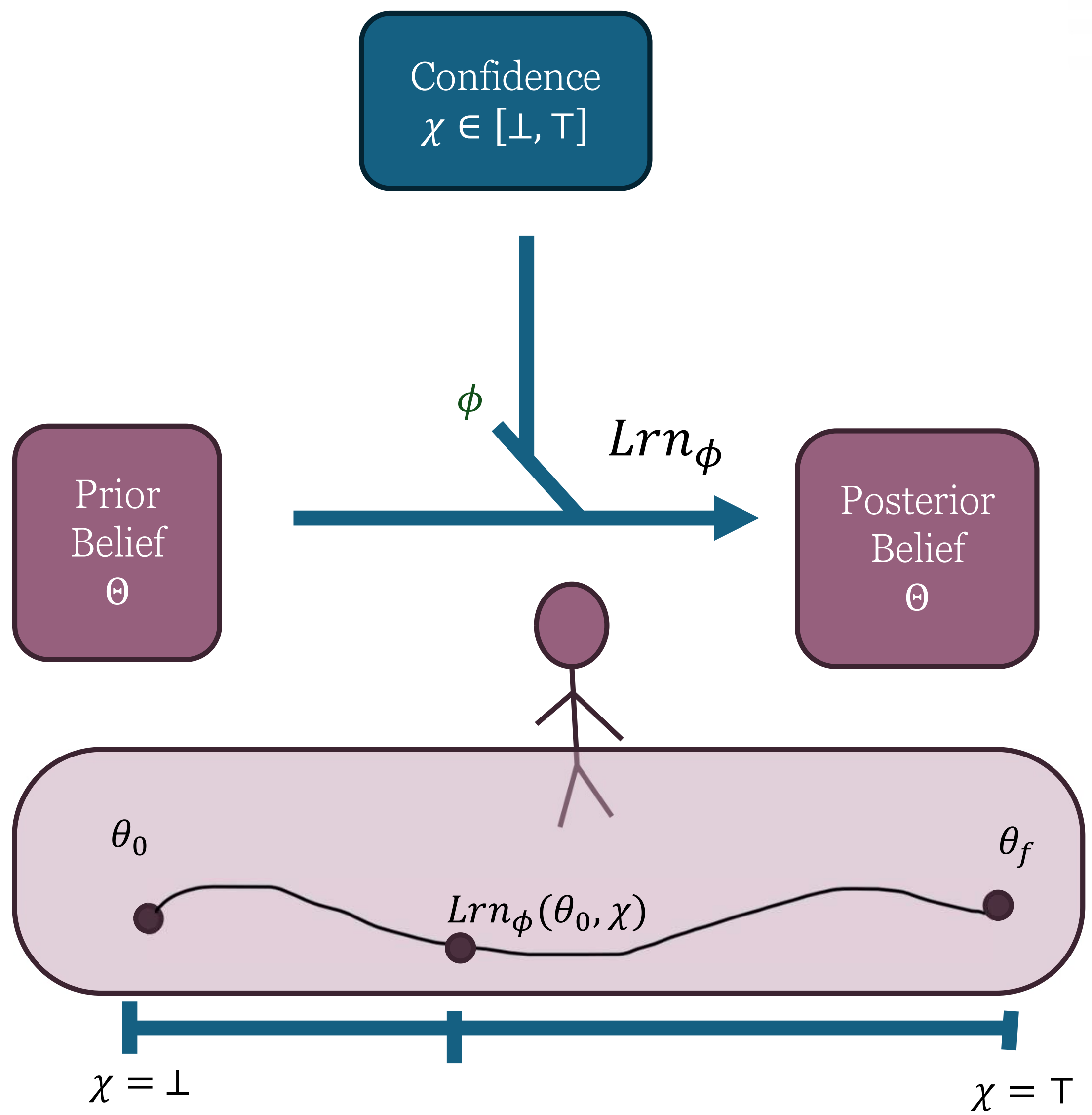
full confidence [FC] $Lrn_\phi^\top \circ Lrn_\phi^\top = Lrn_\phi^\top.$

continuity [L2] $\chi \mapsto Lrn(\theta, \chi, \phi)$
is continuous, twice diffble

residuals [L3] $\chi < \chi' \implies$
$\exists \chi''. \bot < \chi'' \leq \chi'$
$Lrn_\phi^{\chi''} \circ Lrn_\phi^{\chi}(\theta) = Lrn_\phi^{\chi'}(\theta).$

acyclic [L4] If $\chi_0 \leq \chi \leq \chi_1$
and $Lrn_\phi(\chi_0, \theta) = Lrn_\phi(\chi_1, \theta)$,
then $Lrn_\phi(\chi, \theta) = Lrn_\phi(\chi_0, \theta)$.

combinative [L5] $Lrn_\phi(\chi, Lrn_\phi(\chi', \theta))$
$= Lrn_\phi(\chi \oplus \chi', \theta)$



Canonical Representations of Confidence

Theorem (additive representation).
If Lrn satisfies [L1-5], then there is a translation $g(\chi, \theta)$ *of confidence* $\chi \in [\bot, \top]$ *to the additive domain* $[0, \infty]$ *and a learner* ^{+}Lrn *such that*

$$Lrn(\phi, \chi, \theta) = {}^{+}Lrn(\phi, g(\chi, \theta), \theta)$$

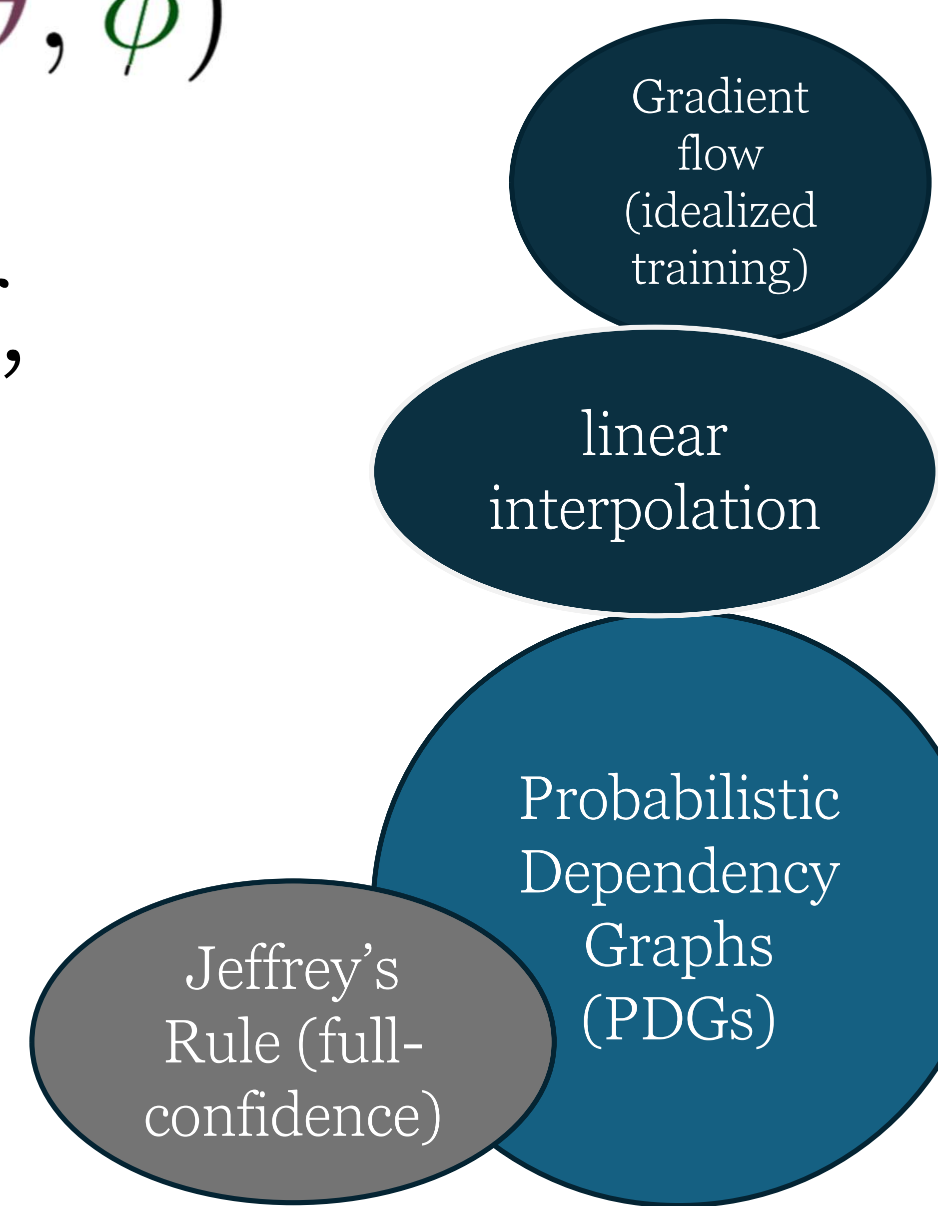
- This "flow form" implies a vector field representations of learners which can be very useful;

Optimizing Learners

[LB4] $$\frac{\partial}{\partial \chi} Lrn(\phi, \chi, \theta) = \nabla_\theta Bel(\theta, \phi)$$

learning is about locally increasing belief,
i.e., gradient descent to minimize loss.

Some examples using relative
entropy and log probability:



What about when learning objective is linear?

Defn (Loss-Linear Learner).
An optimizing learner with a linear objective,
i.e., satisfying LB4 with $Bel(\theta, \phi) = \mathbb{E}_\theta\left[V_\phi \right]$,
in the natural (Fisher) geometry.

Defn (Bayesian Learner).

- Beliefs correspond to $P(H)$;
- H comes with likelihood $P(\phi \mid H)$;
- Updates by Bayes Rule: $\exists \star \in [\bot, \top]$.
  $Lrn\big(\phi, \star, P(H)\big) = P(H \mid \phi) \propto P(\phi \mid H) P(H)$

Proposition. The additive form of a loss-linear learner is:

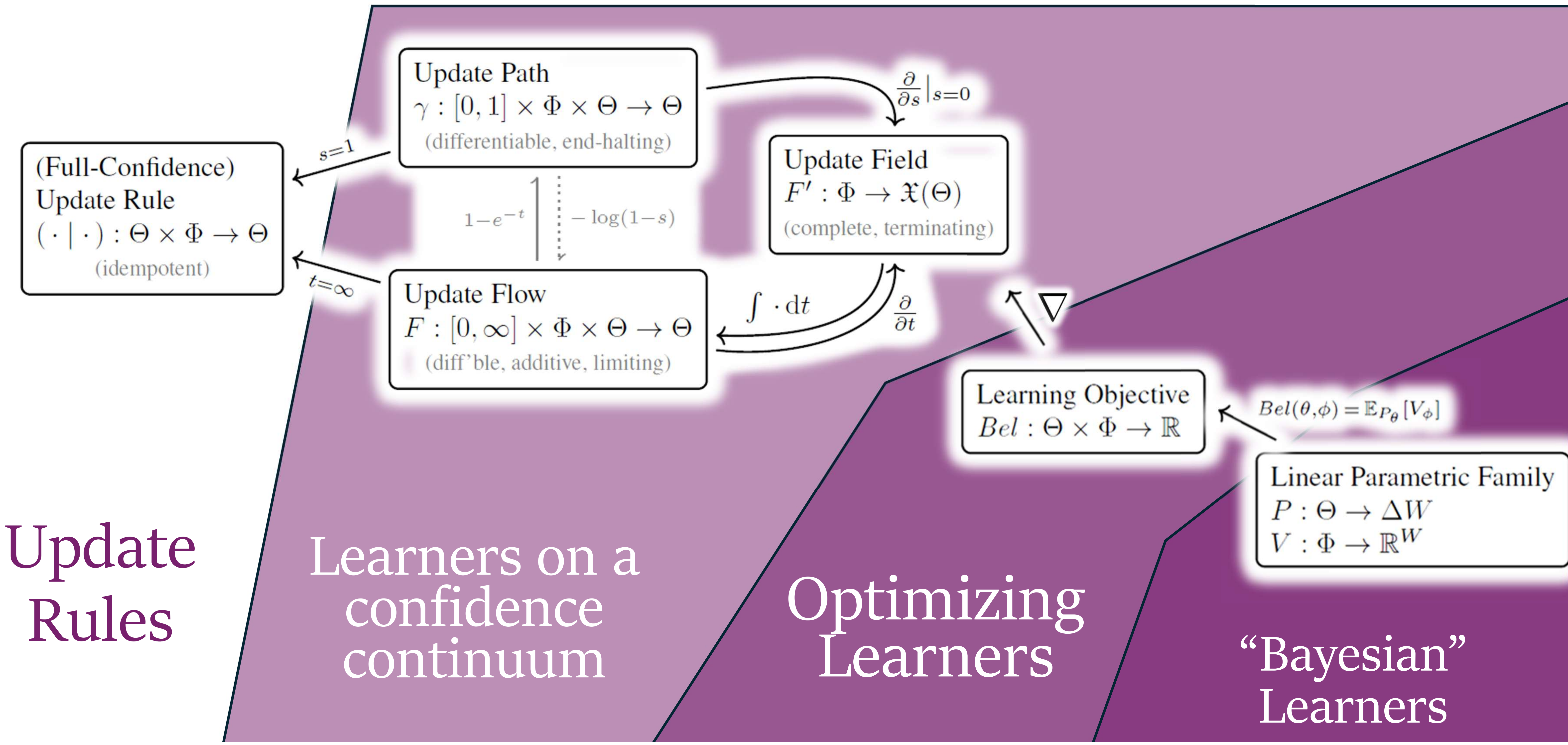
$$Boltz(P, \beta, \phi)(w) \propto P(w) \exp\Big(\beta\, V_\phi(w) \Big).$$

That is, the posterior is a Boltzman distribution with the prior as the base measure, the confidence as inverse temperature, and the value V_ϕ as the energy.

Proposition: A learner for probability distributions is Bayesian if and only if it is loss-linear, with

$$V_E(h) = \log P(E|h)$$

Representations of Confidence-based Learners



Conclusion

If certainty is about black and white,
then probability is about shades of gray,
learner's confidence is about transparency.

- Learner's confidence is distinct from likelihood;
- Unifies many concepts in the literature:
  - Sensor precision, Kalman gain, virtual evidence, weight of evidence, thermodynamic coldness, Boltzmann rationality constant β, learning rate, number of epochs, ...
- Bayesian updates are a restrictive special case.