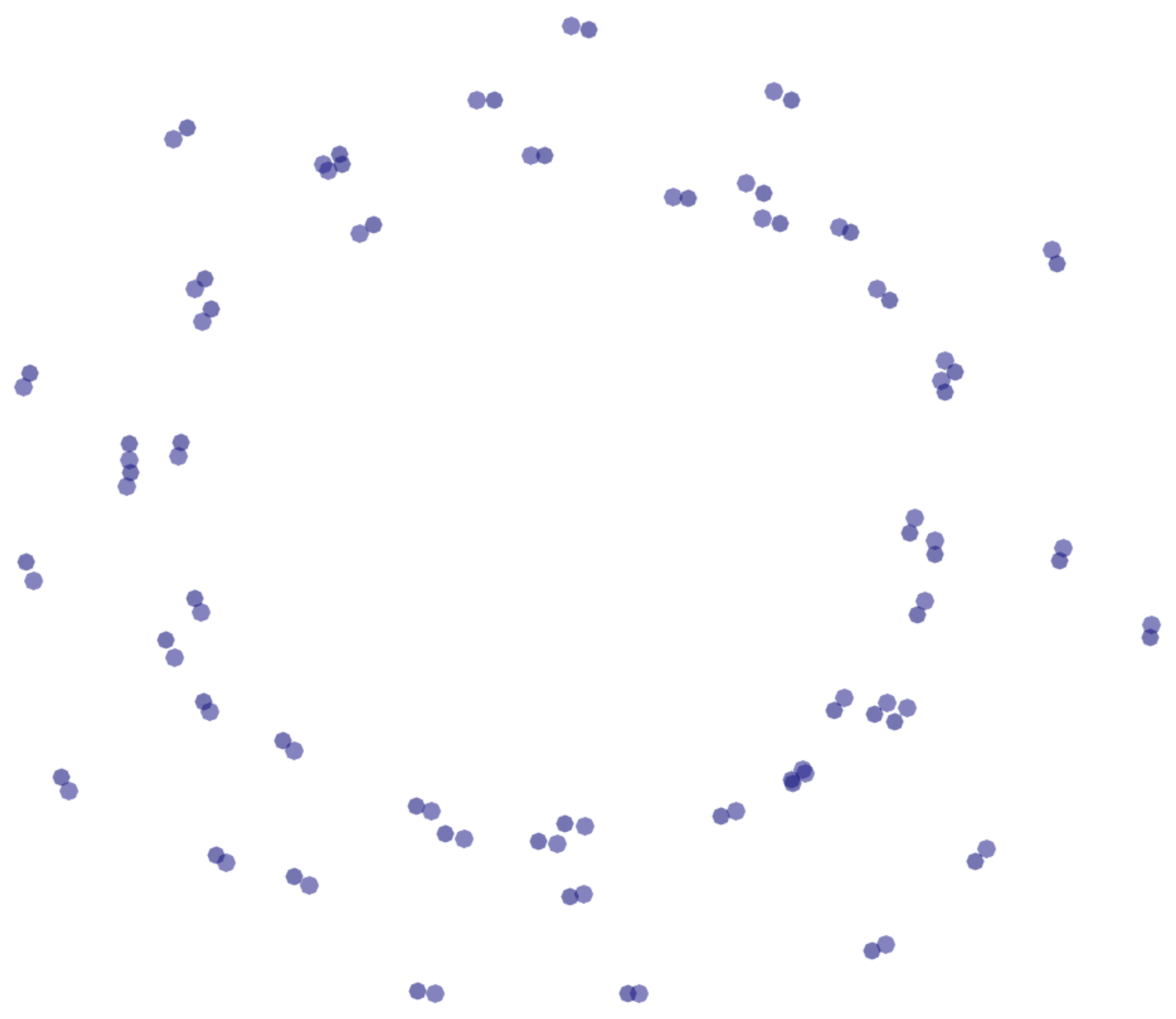
Collective behavior from surprise minimization



Conor Heins

@conorheins

VERSES Machine Learning Lab

Max Planck Institute of Animal Behavior

Modelling Multi-Scale Collective Intelligences Workshop

19 November 2024

Overview

Part I: Background on Bayesian cognitive science & active inference

- Perception as unconscious inference and Bayesian inference
- Minimizing prediction error as an algorithm for inference
- Active inference

Part II: Applying the concepts from Part I to model collective motion

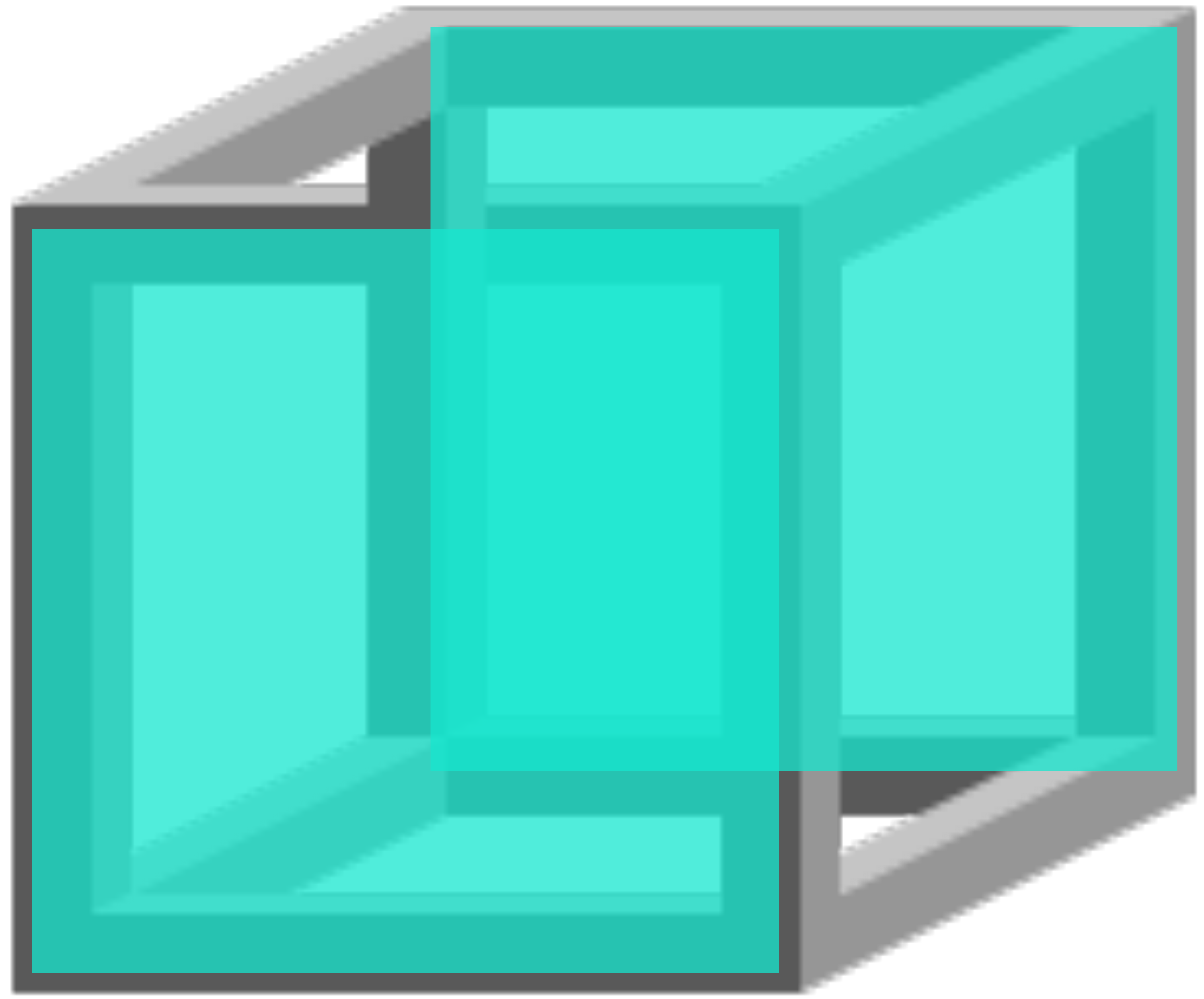
- Collective motion overview
- Phenomenological models vs ‘cognitive' models
- Collective motion from multi-agent active inference
- Emergent information transfer & decision-making
- Learning one’s model online, i.e., ‘behavioral plasticity’

The "Bayesian turn" in the cognitive sciences

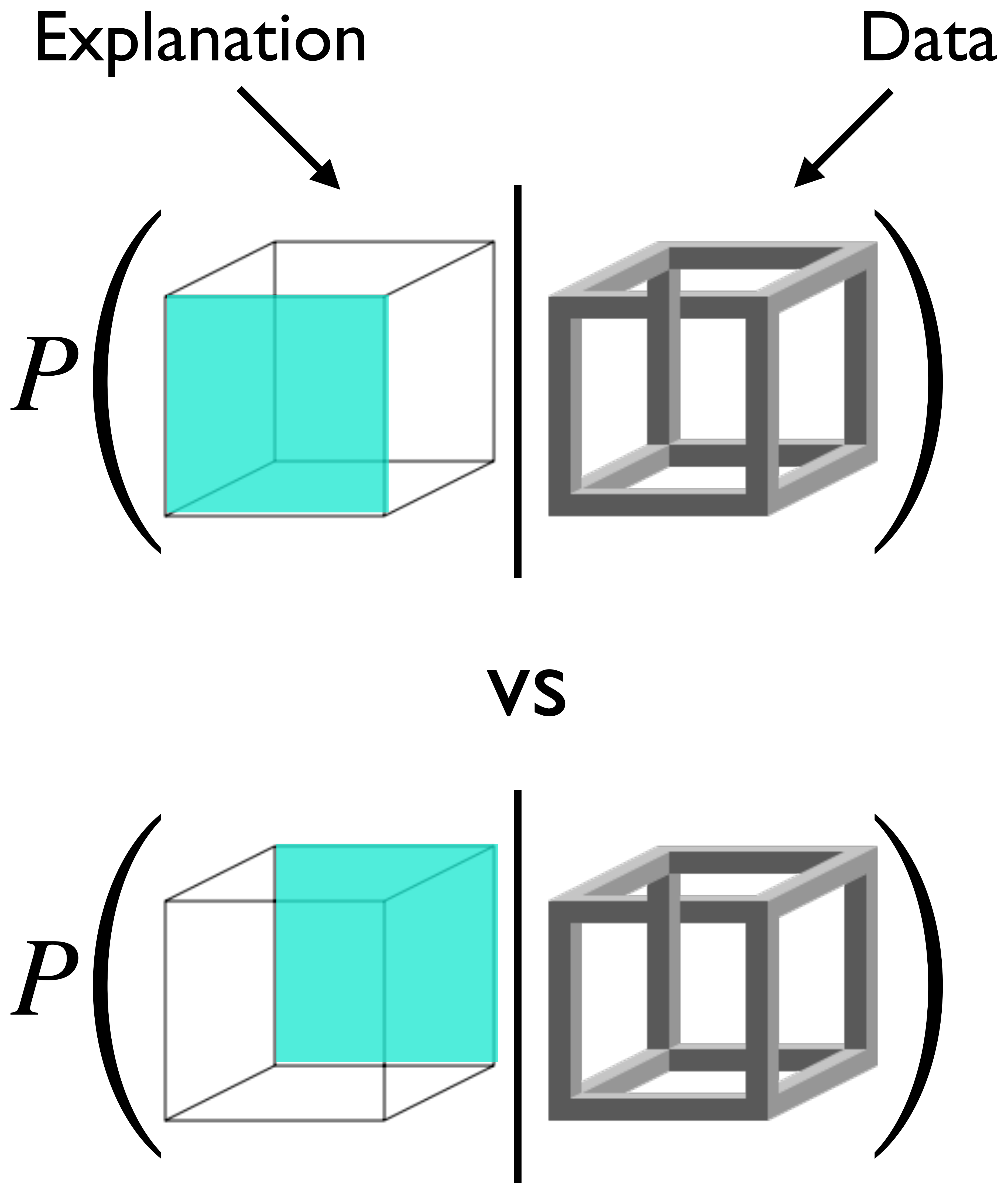
- Hermann von Helmholtz, "Perception as unconscious inference" (*unbewusster Schluss*)

Hermann von Helmholtz

Perception as explanation

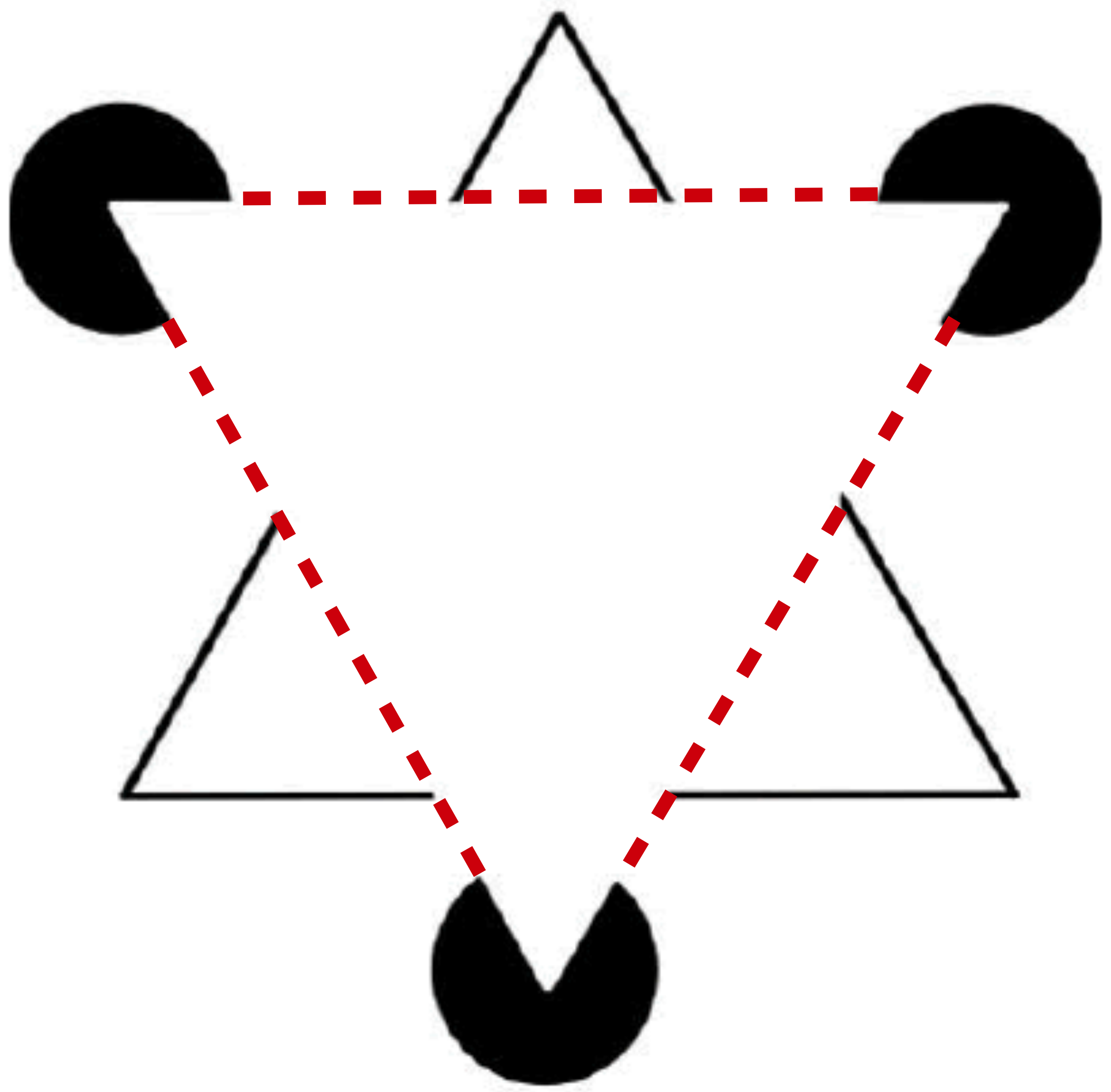


Neckar Cube



Perception as explanation

Kanisza’s triangle



Your perception is not the “raw data” (aka the pixel intensities on the screen), but an **inference** or **interpretation** of that data

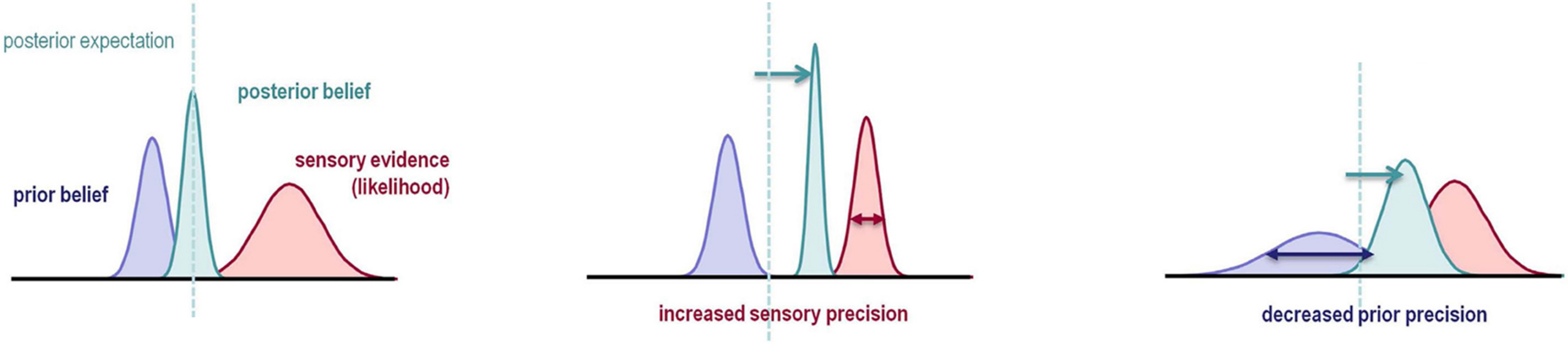
The "Bayesian turn" in the cognitive sciences

- Hermann von Helmholtz, "Perception as unconscious inference" (*unbewusster Schluss*)

Hermann von Helmholtz

- Formalized later in the 20th century as probabilistic inference — use Bayes Rule to compute posterior probabilities

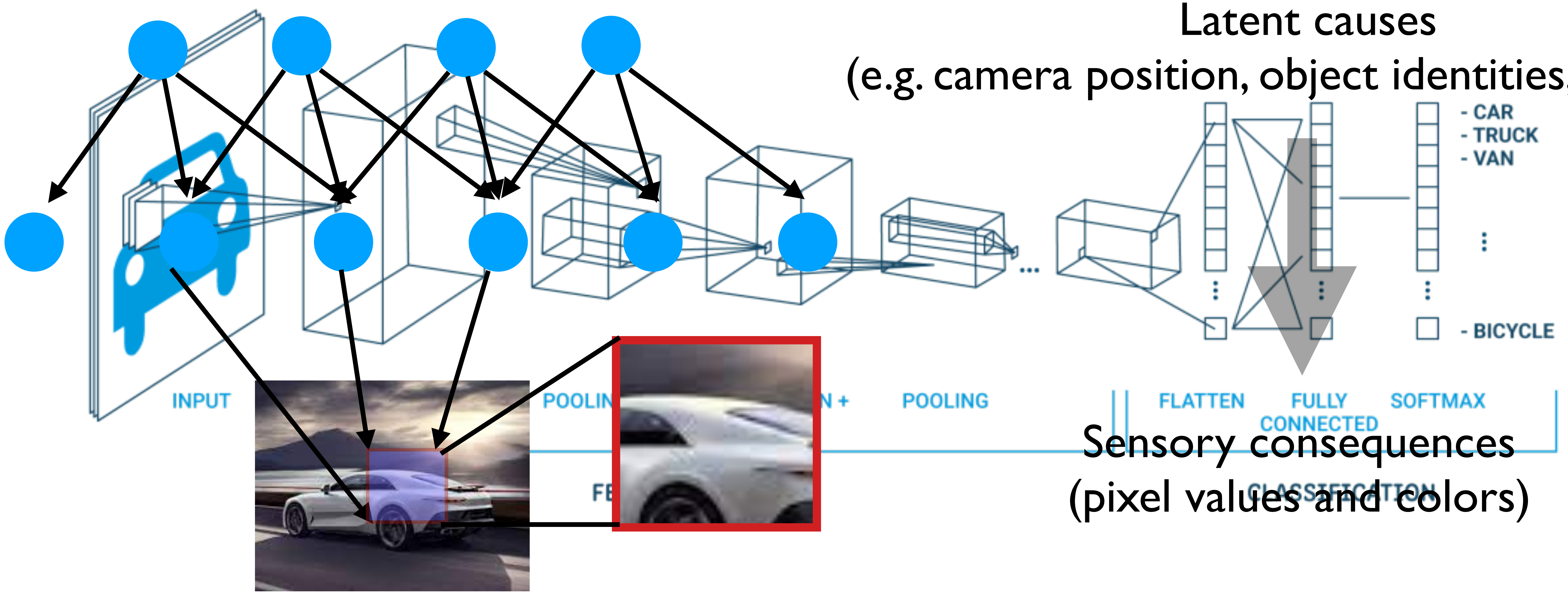
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$



Perception as feature detection

Perception as explanation

Latent causes
(e.g. camera position, object identities, locations)

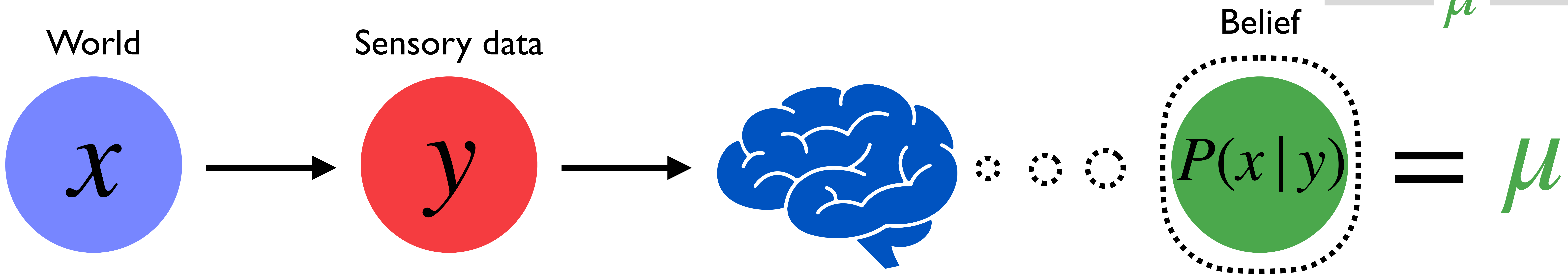


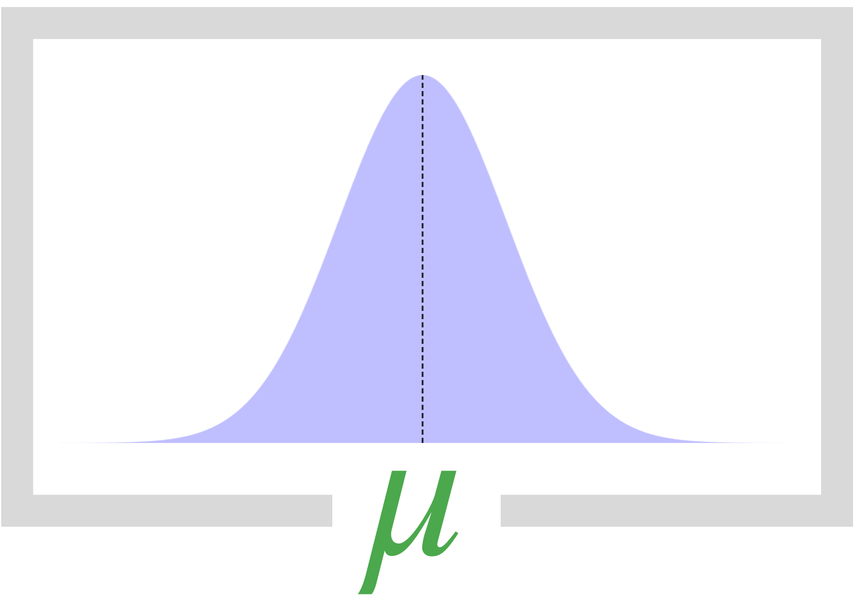
Sensory consequences
(pixel values and colors)

Feedforward cascade of feature detectors

Inference using a generative model (inverse graphics)

One way to infer:
minimise **prediction error**



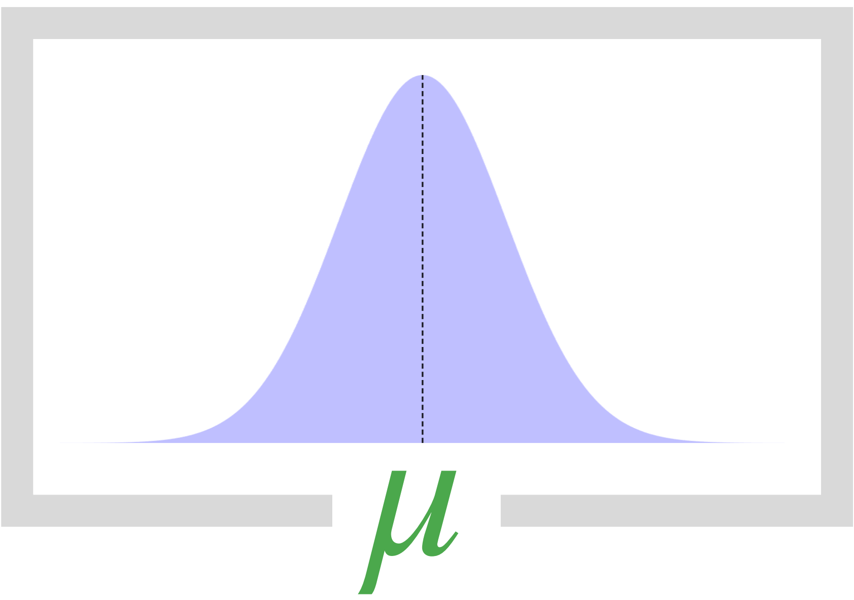


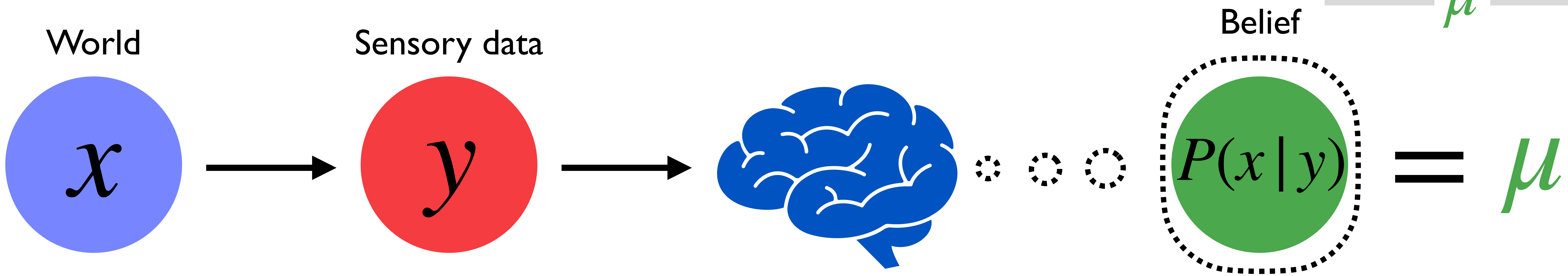
$$\mu^* = \underset{\mu}{\arg\min}\left(\text{Prediction error}\right)$$

What you see What you expect

More formally:

use a gradient descent on free energy
to update the posterior means





$$\frac{d\mu}{dt} = -\frac{\partial F(\mu, y)}{\partial \mu} \longrightarrow F \propto \epsilon_y^2 + \epsilon_\mu^2$$

Sensory prediction errors

$\epsilon_y = y - \mu$

Prior prediction errors

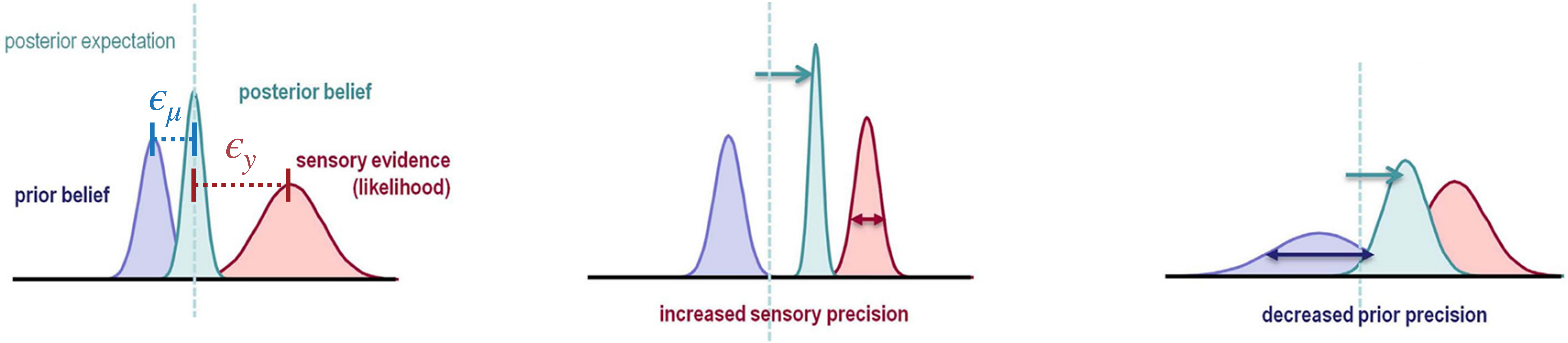
$\epsilon_\mu = \mu - \eta$

Precision weighting is important!

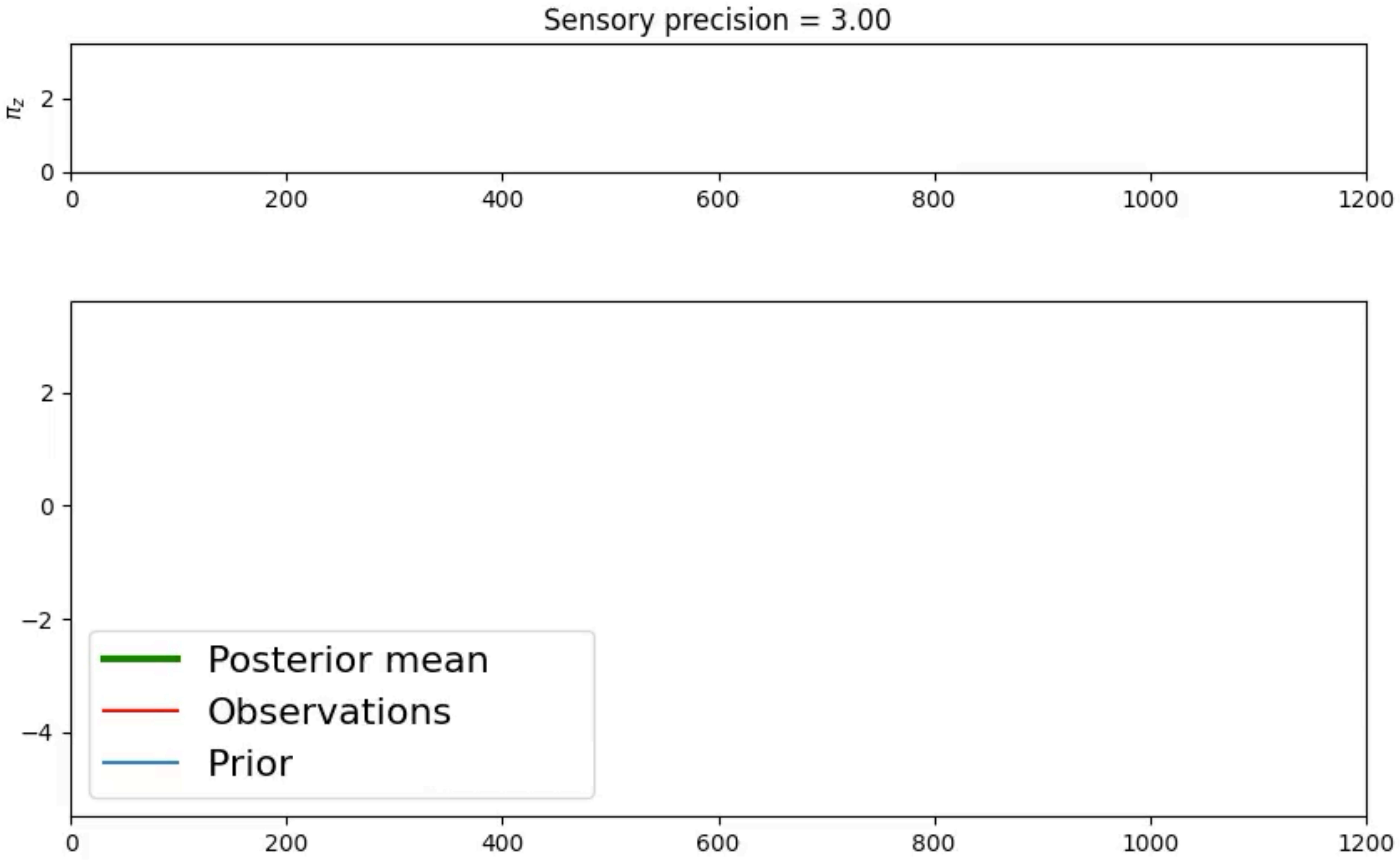
$$\frac{d\mu}{dt} = -\frac{\partial F(\mu, y)}{\partial \mu}$$

$$F \propto \pi\epsilon_y^2 + \pi\epsilon_\mu^2$$

Free energy
= surprise
= "sum of (squared) *precision-weighted prediction errors*"

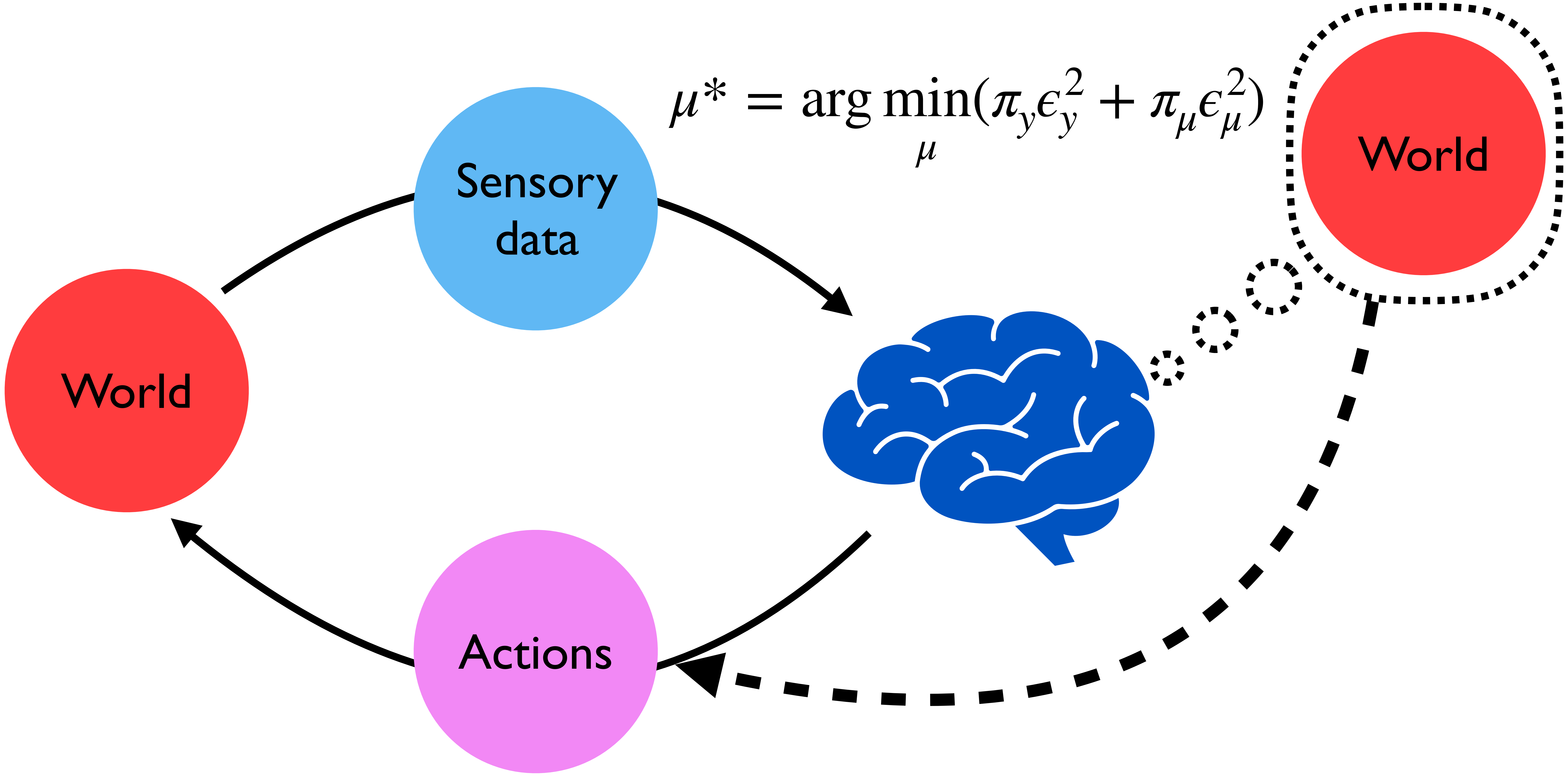


Bayesian filtering example with attenuation in sensory precision



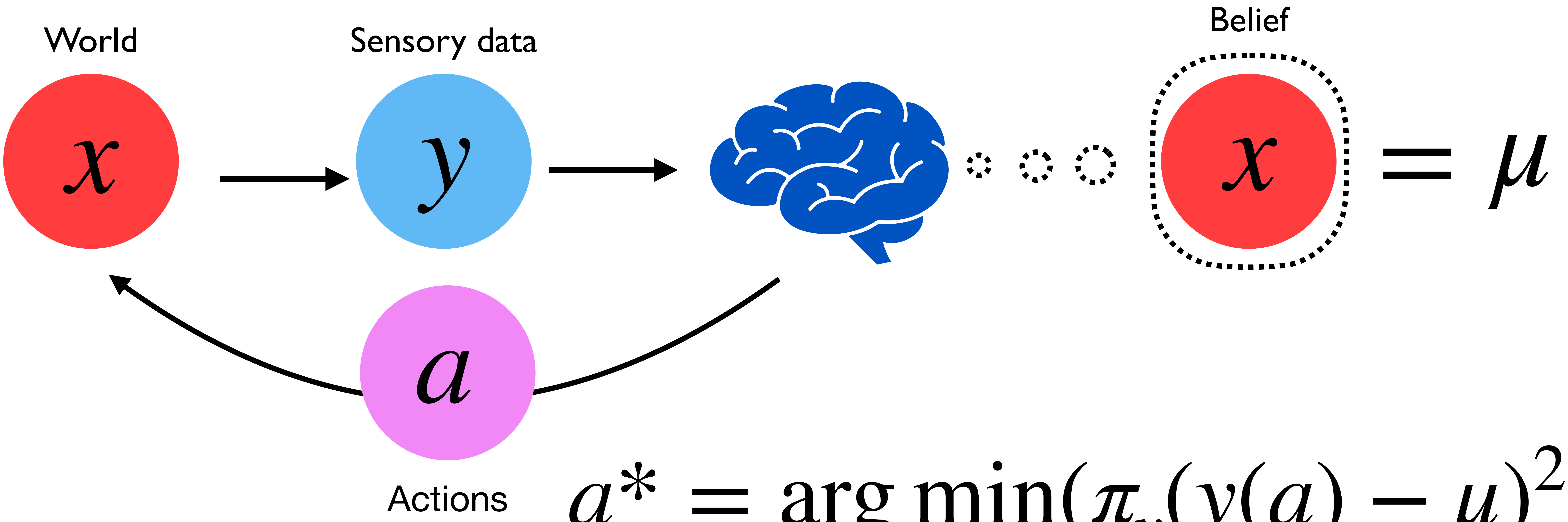
What about action?

Active inference, aka belief-guided action



This is a standard reinforcement learning formulation
— how is active inference any different?

How to act?
also minimise **prediction error**



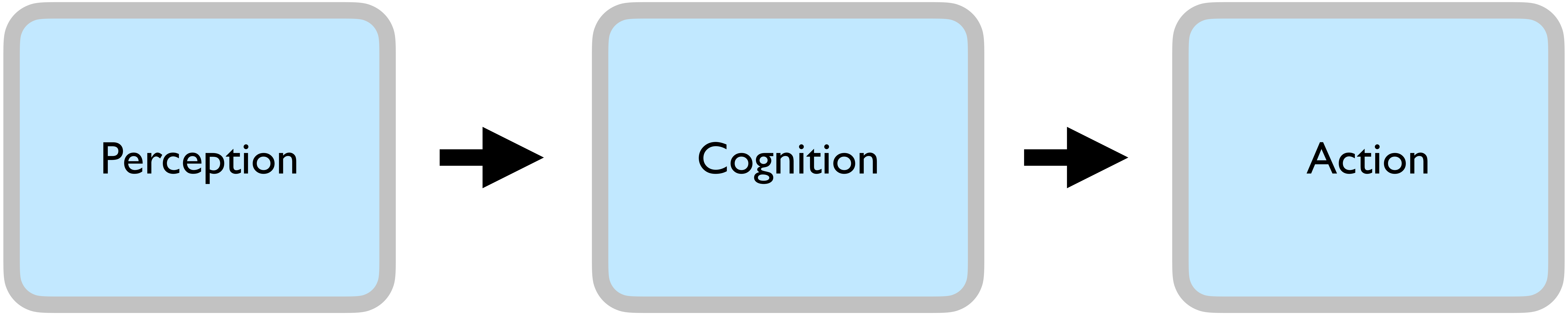
$$a^* = \arg\min_a(\pi_y(y(a) - \mu)^2)$$

Under **active inference**, everything is about minimising **prediction error** aka "surprise"

There are two ways to become less surprised

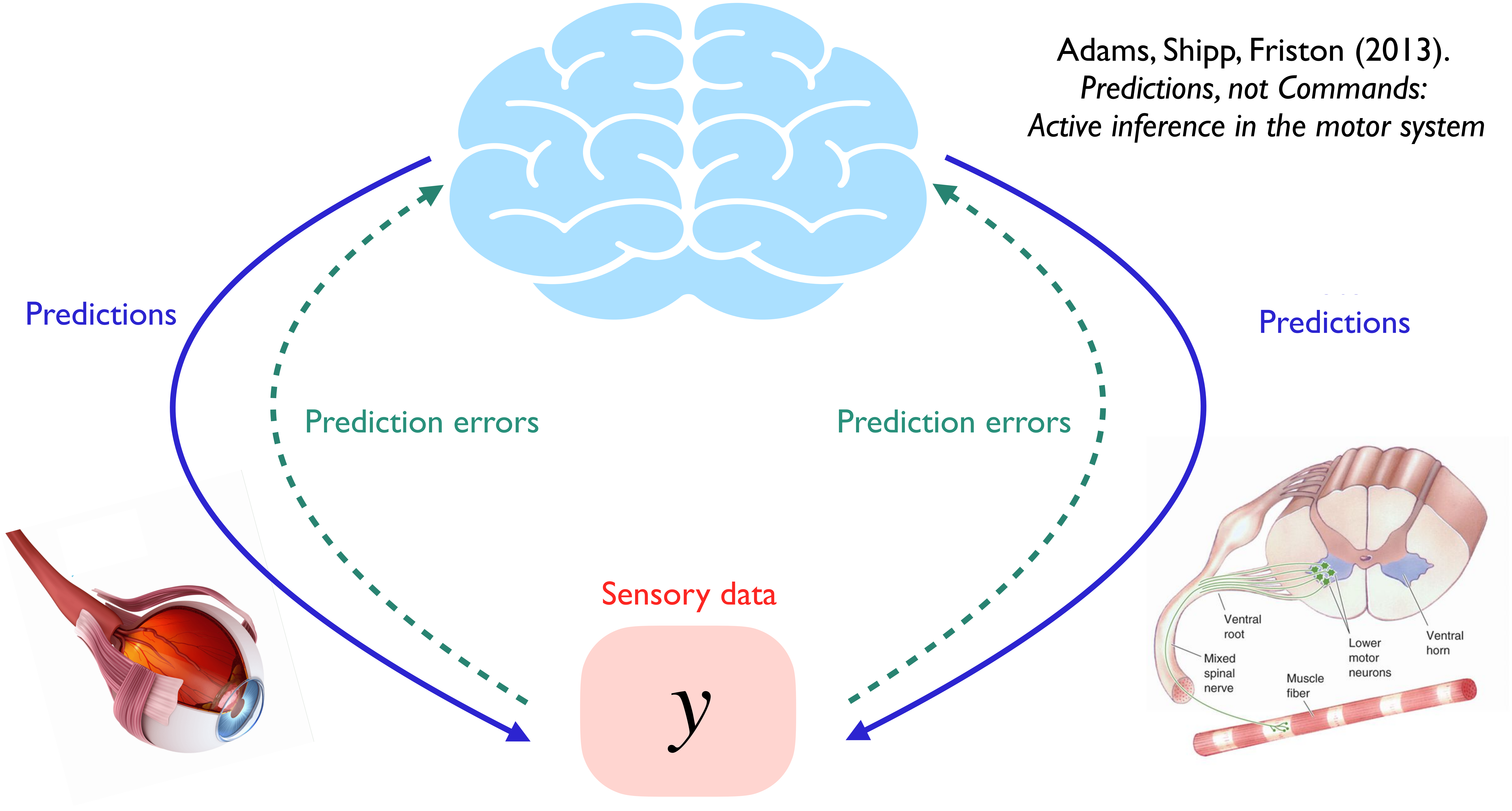
Classical "Sandwich Model" of Cognition

- Unidirectional information flow from sensory states to motor effectors
- Distinct stages of processing
- Cognition plays role of generating "actionable representations"



Susan Hurley, *Synthese* 2001

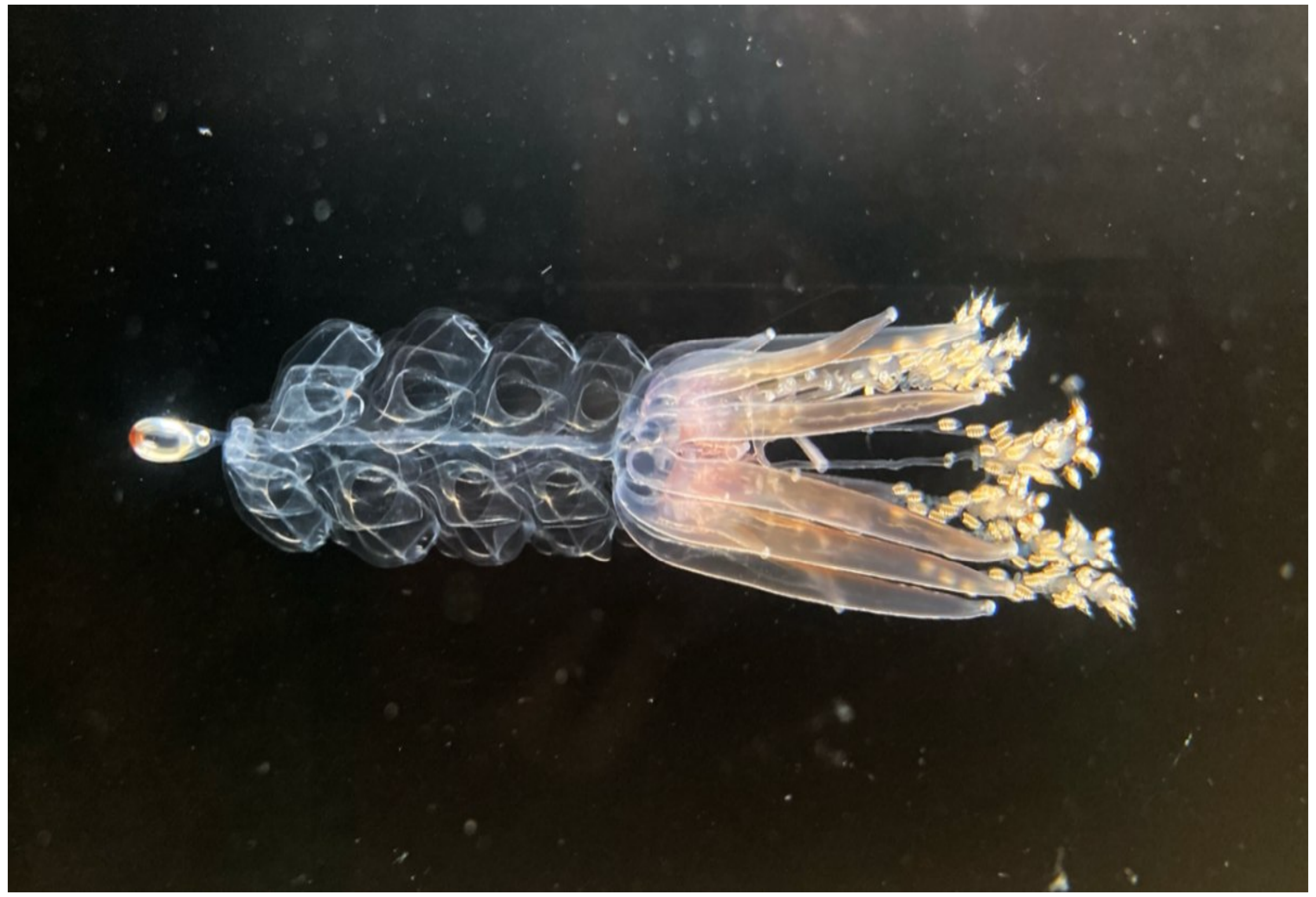
Adams, Shipp, Friston (2013).
Predictions, not Commands:
Active inference in the motor system

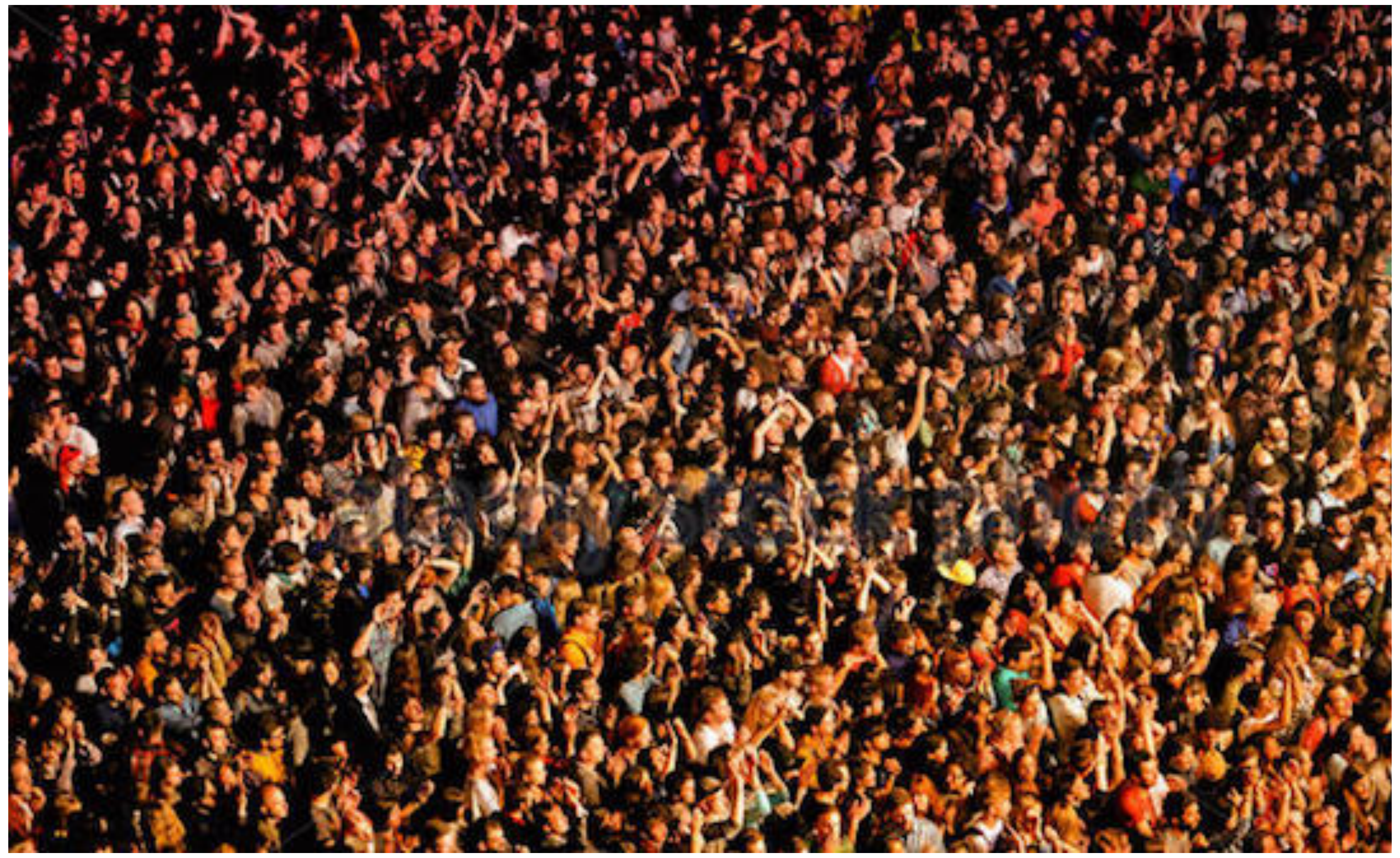


Applications of active inference

- Eye movements and visual foraging (Friston et al. 2012, Mirza et al. 2016, Parr & Friston 2018, …)
- Kinematic and postural control (Maselli et al. 2022, Priorelli et al. 2023)
- Embodied spatial decision-making (Priorelli et al. 2024)
- Emotion recognition (Smith et al. 2019, Hesp et al. 2021, Mirza et al. 2021)
- Economic decision-making under uncertainty (Smith et al. 2020, Markovic et al. 2021)
- Language understanding and speech (Parr & Pezzulo 2021, Friston et al. 2020)
- Active sensing — e.g., whisking in rodents (Mannella et al. 2021)

Collective motion in nature

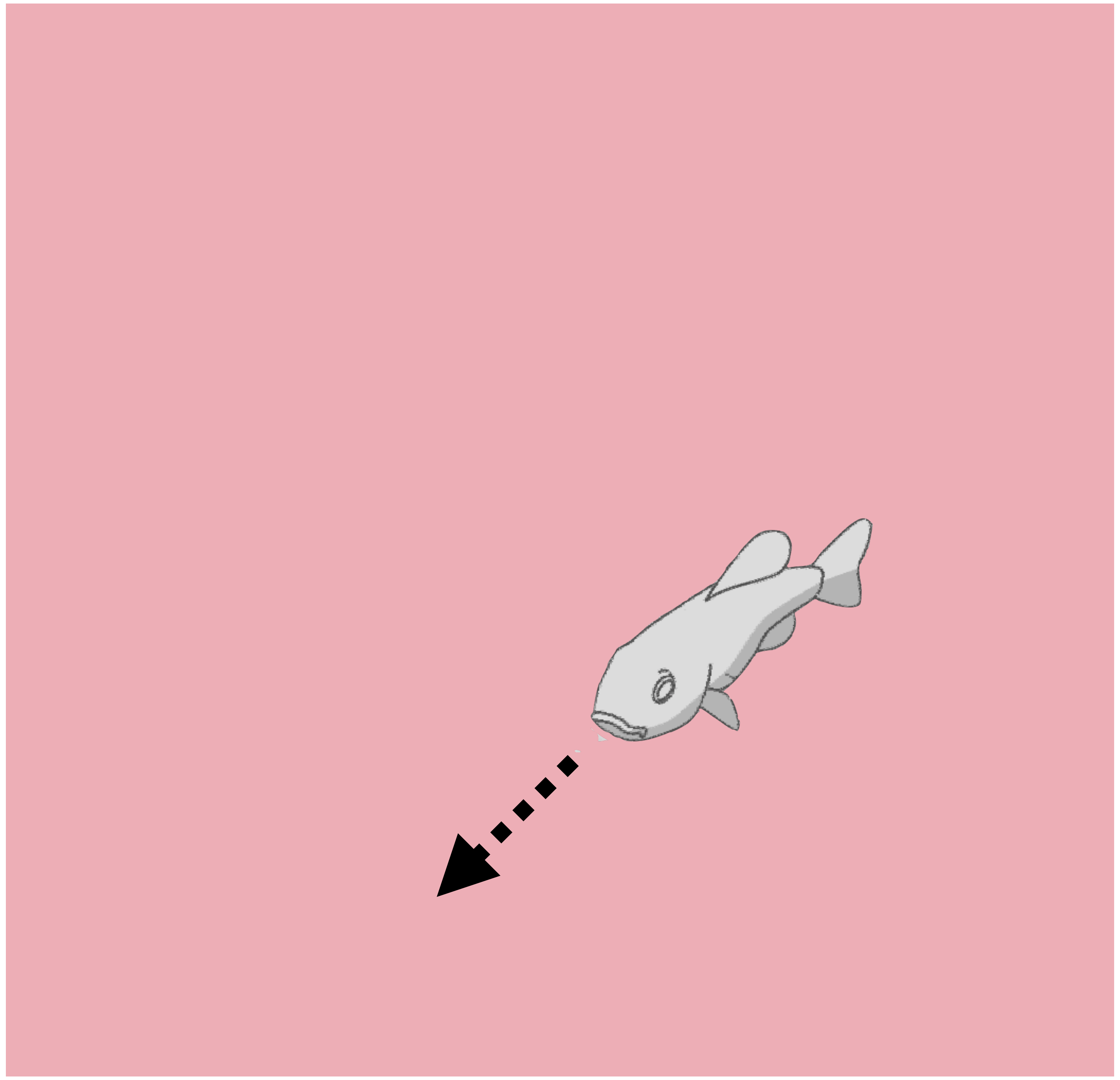




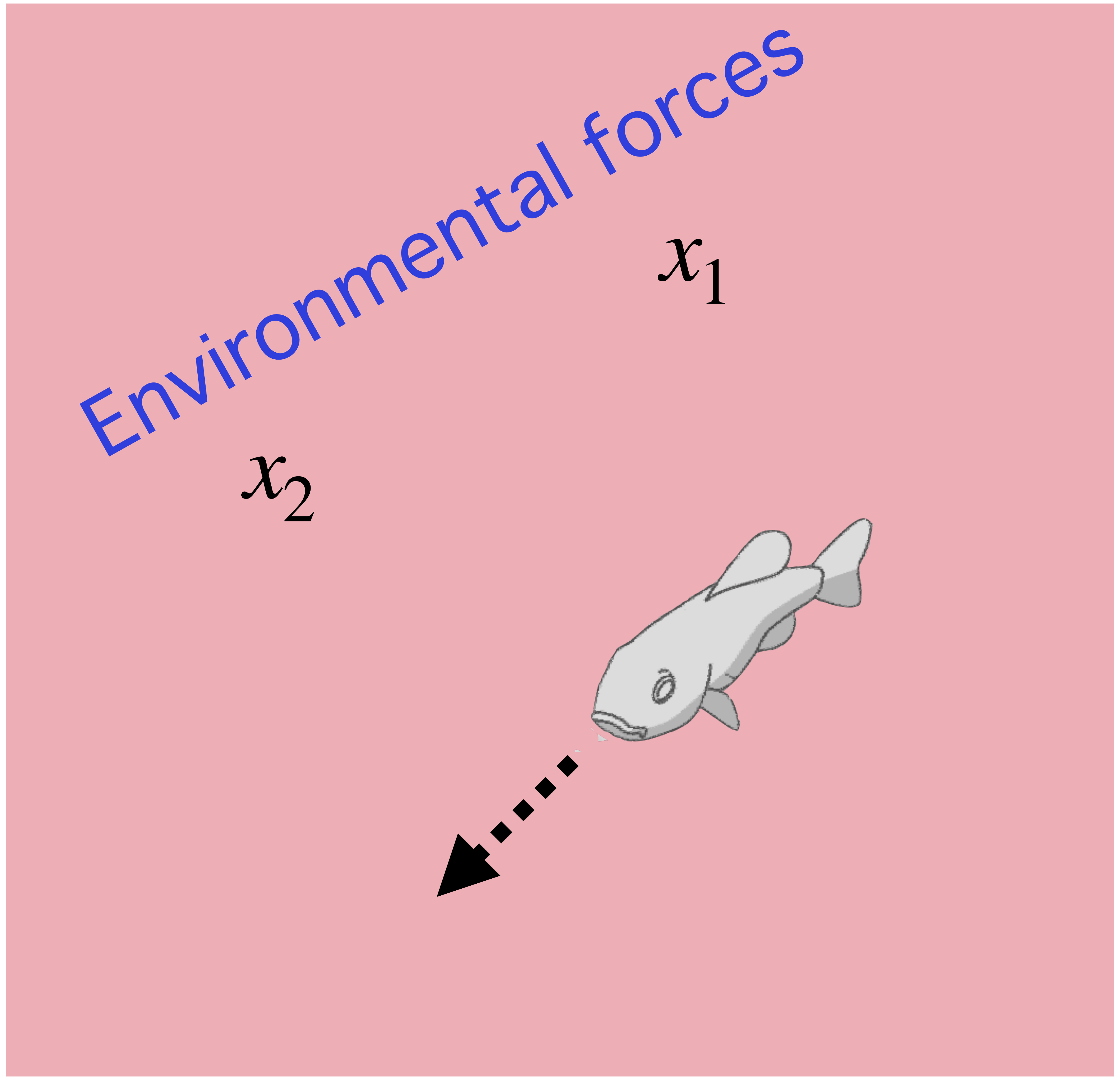
Collective motion in nature

- Example of emergent order from simple, decentralised interactions
- Universality? Shows up across natural, technological, artificial disciplines
- Collective motion is relevant to biologists for many reasons (evolutionary, ecological, cognitive, neurobiological)

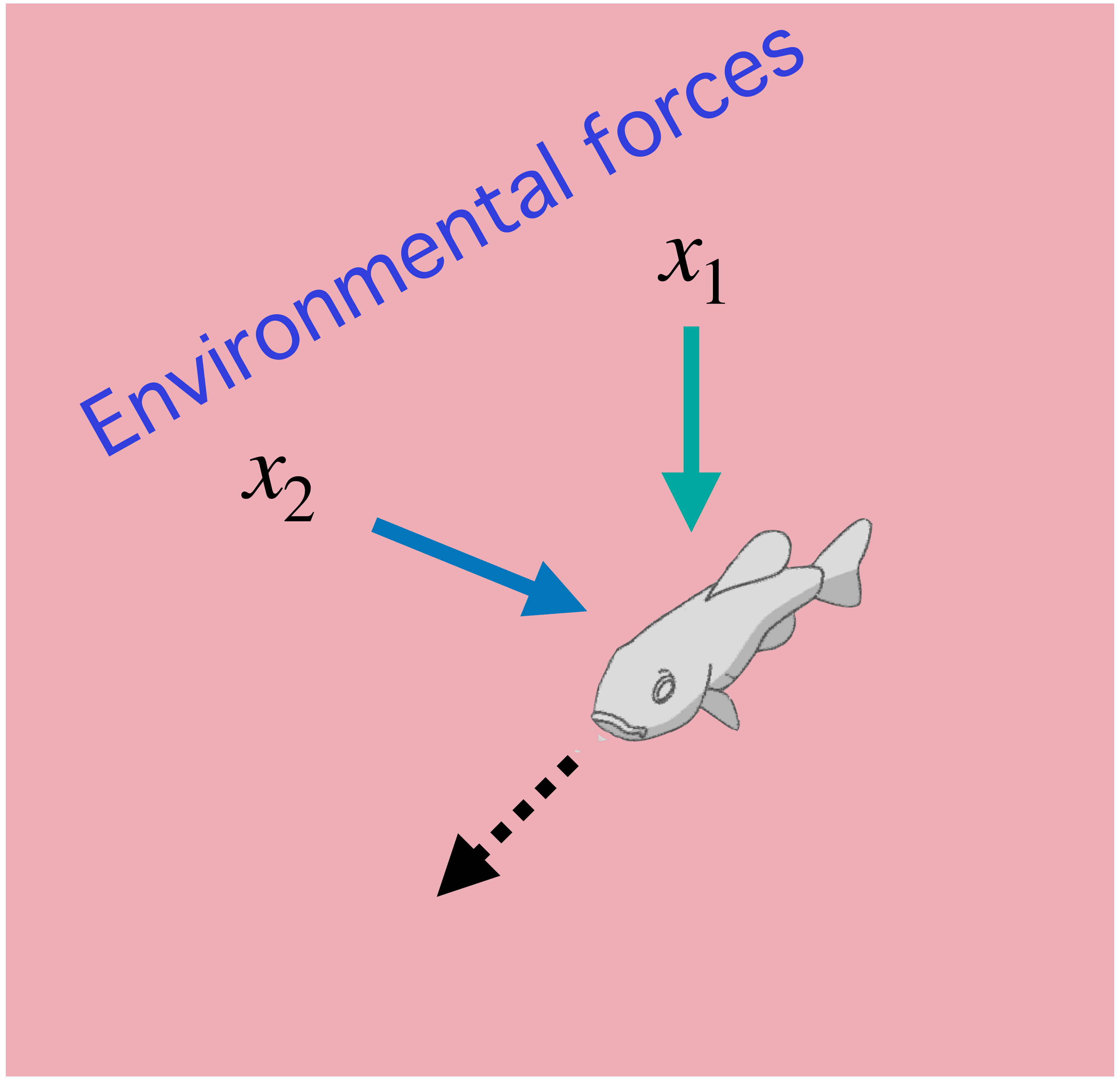
Physics-based models (Vicsek model, etc.)



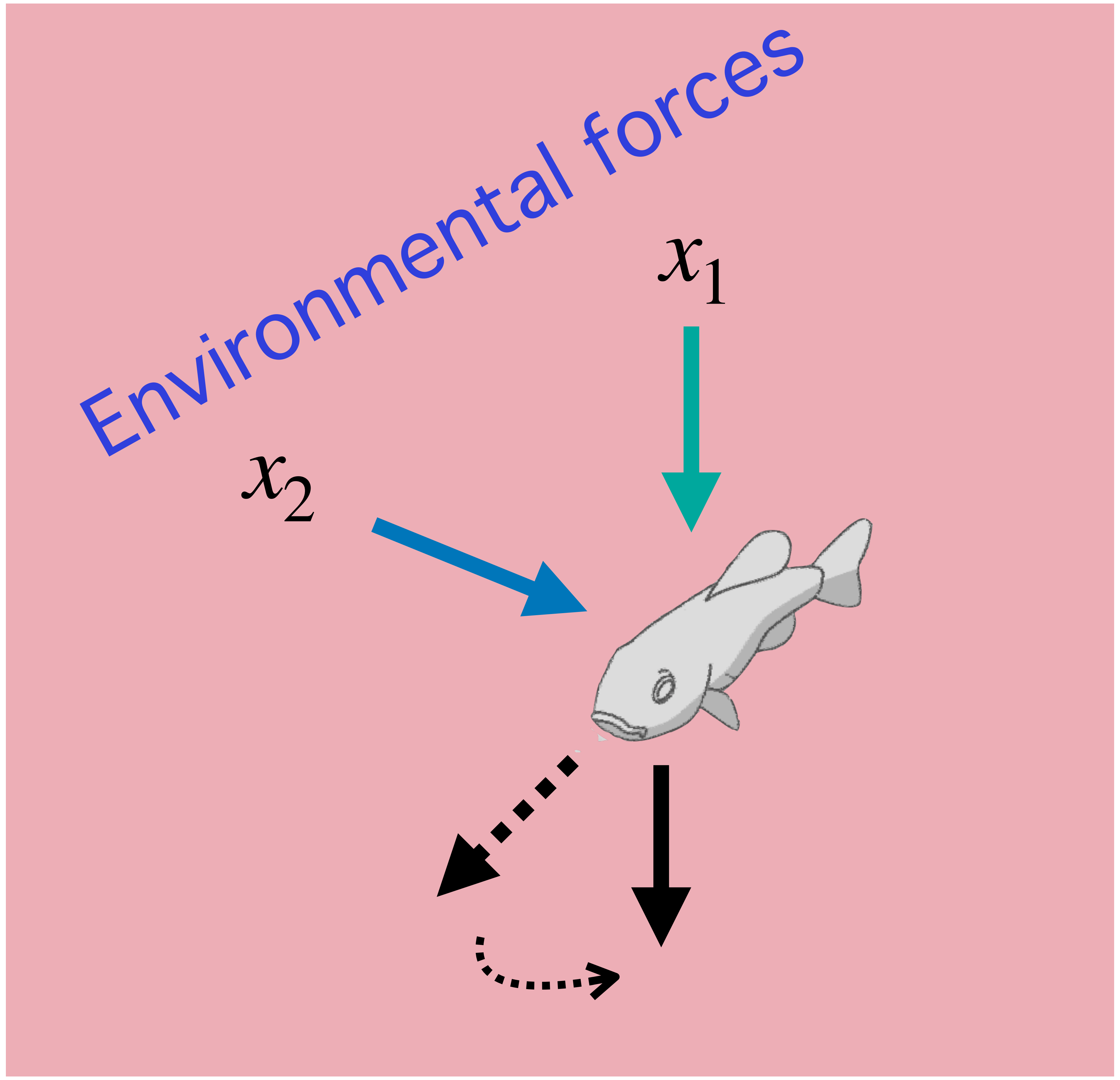
Physics-based models (Vicsek model, etc.)



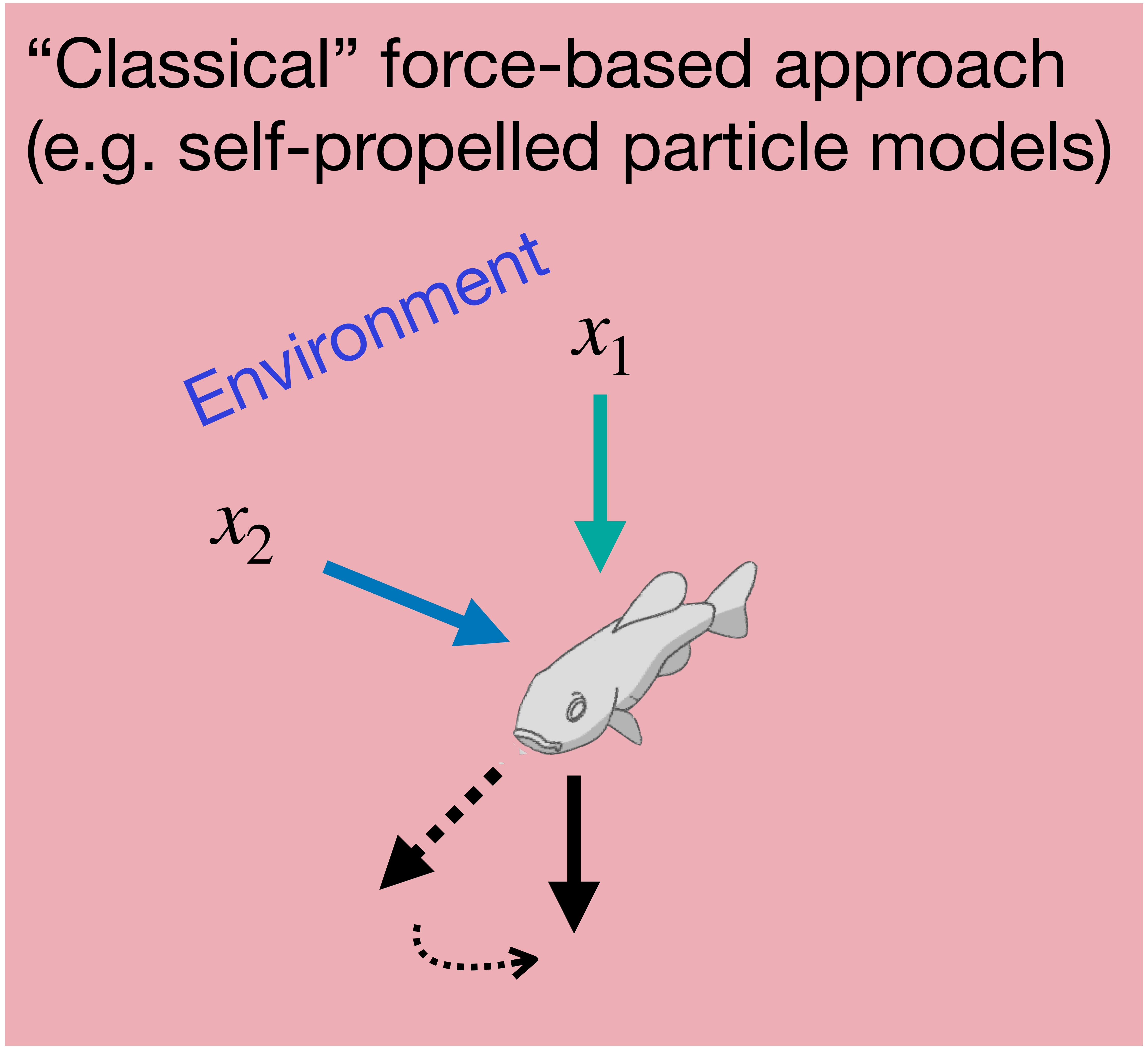
Physics-based models (Vicsek model, etc.)

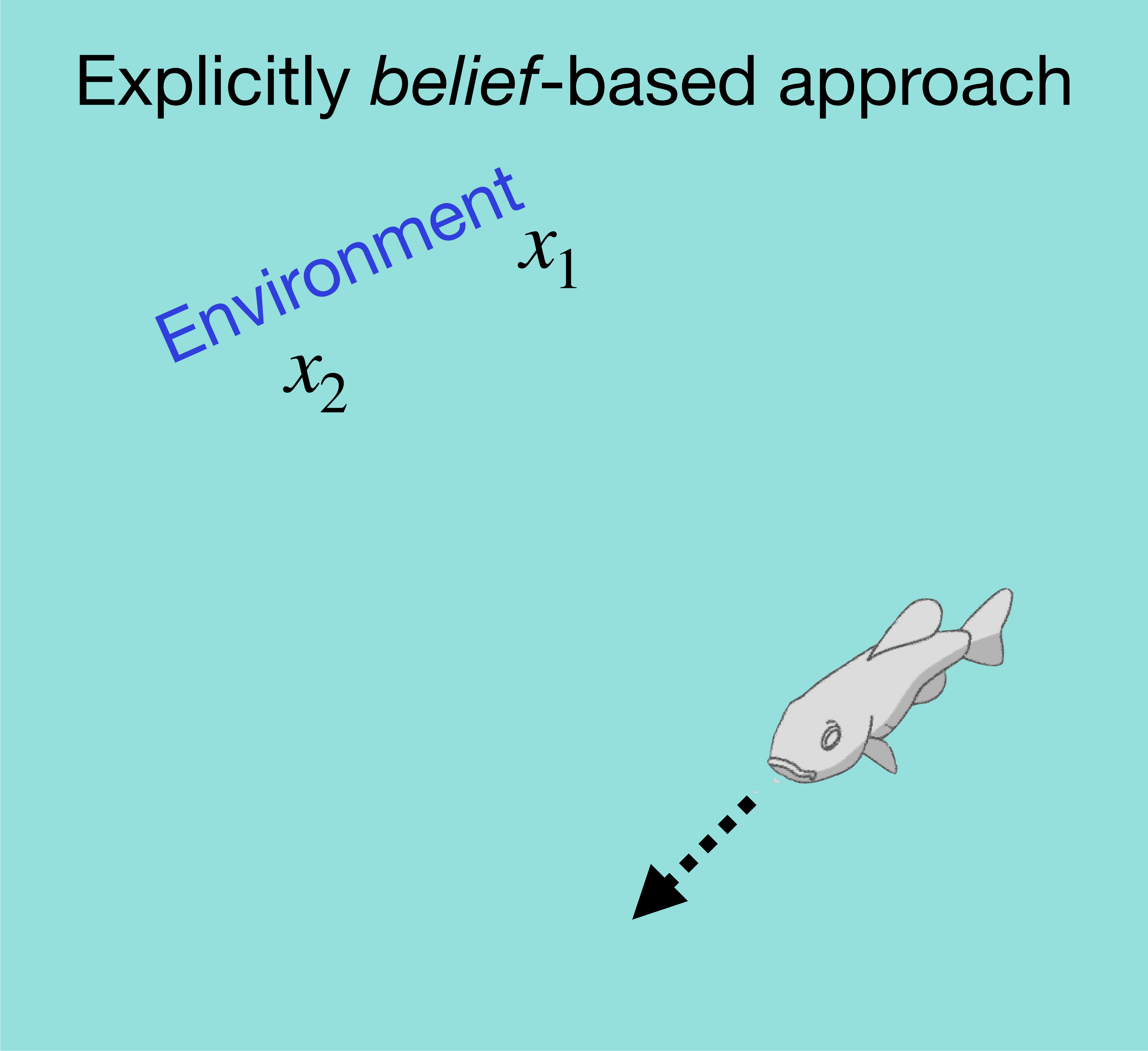


Physics-based models (Vicsek model, etc.)



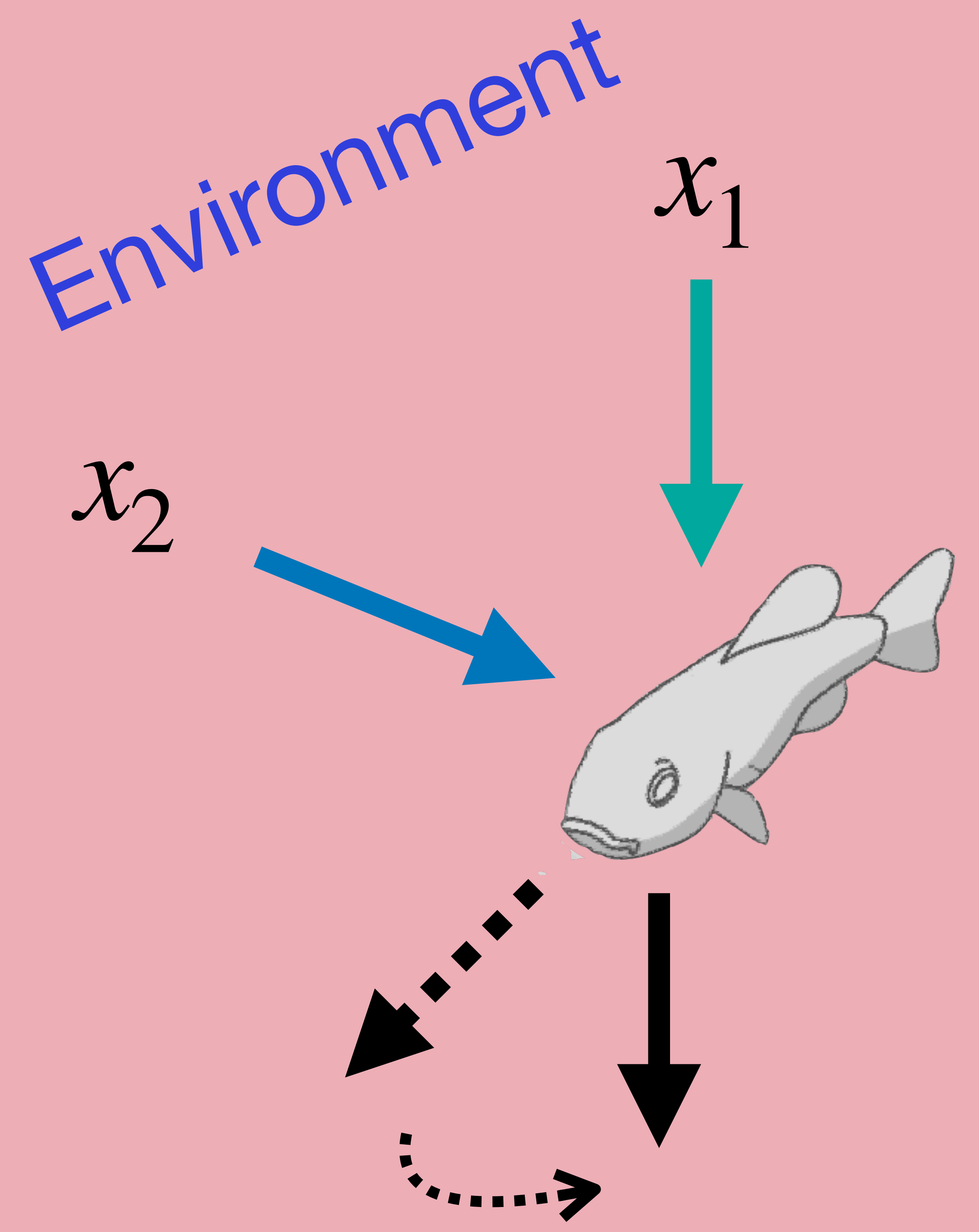
Comparing classical and Bayesian approaches



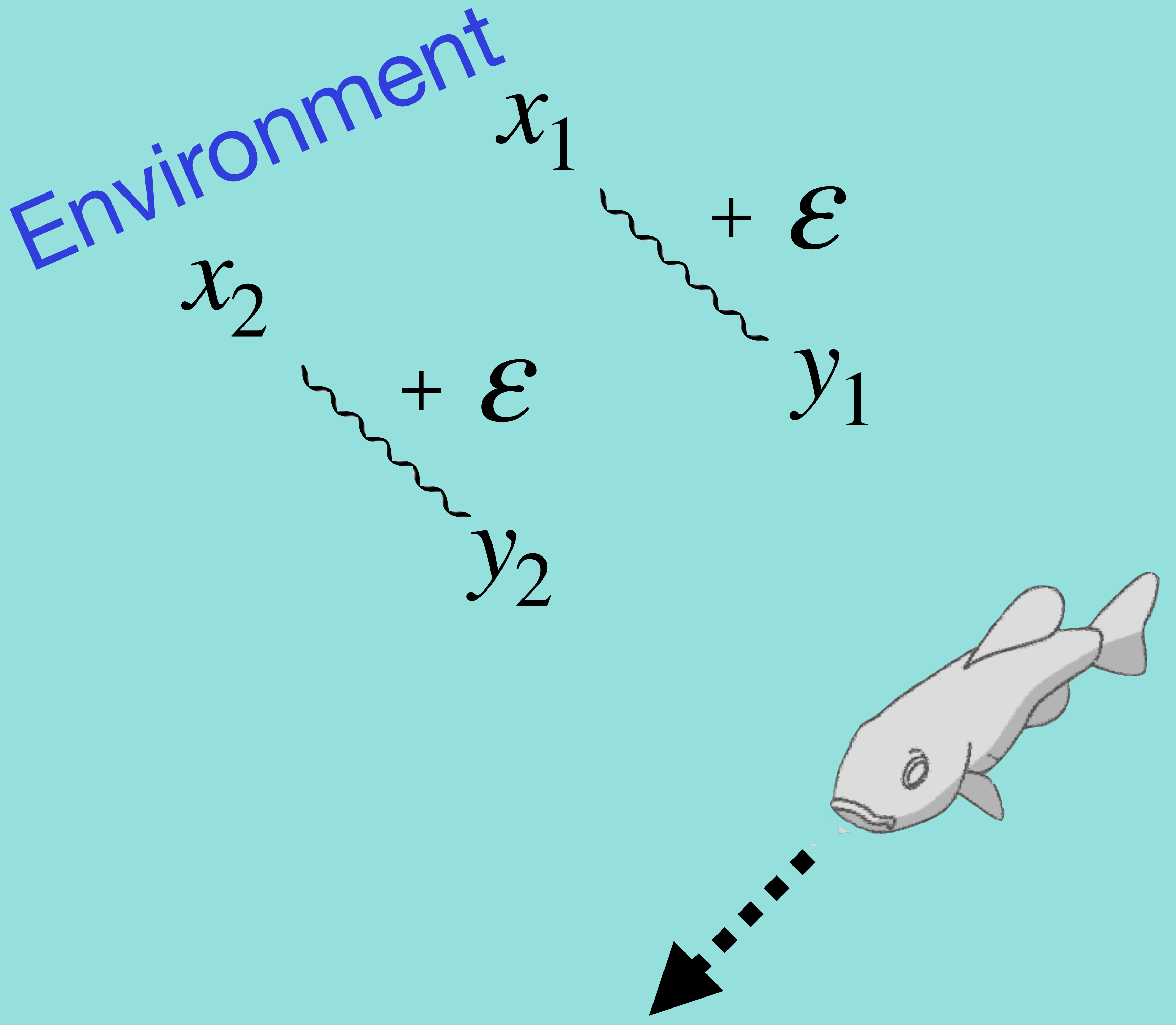


Comparing classical and Bayesian approaches

“Classical” force-based approach
(e.g. self-propelled particle models)

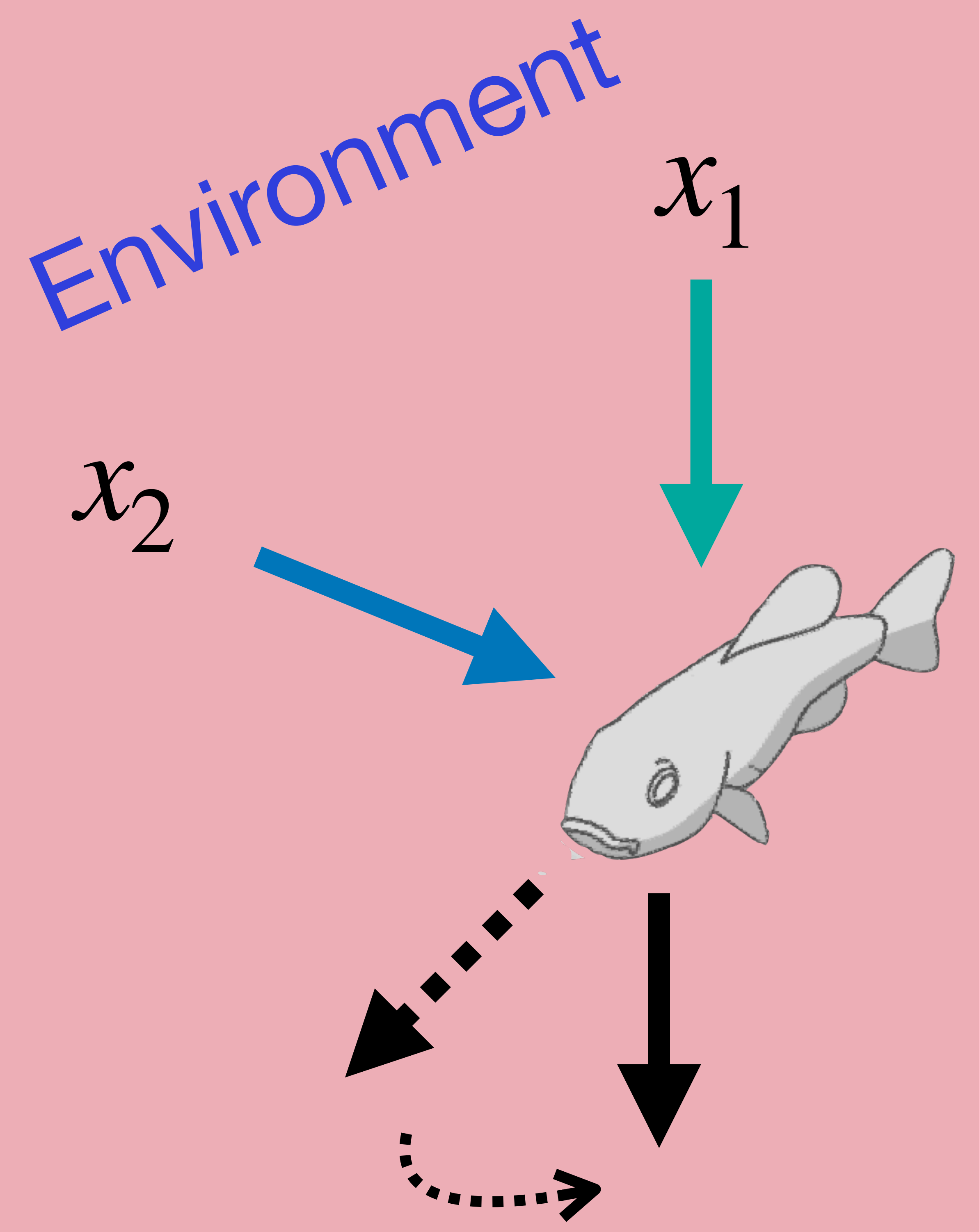


Explicitly *belief*-based approach

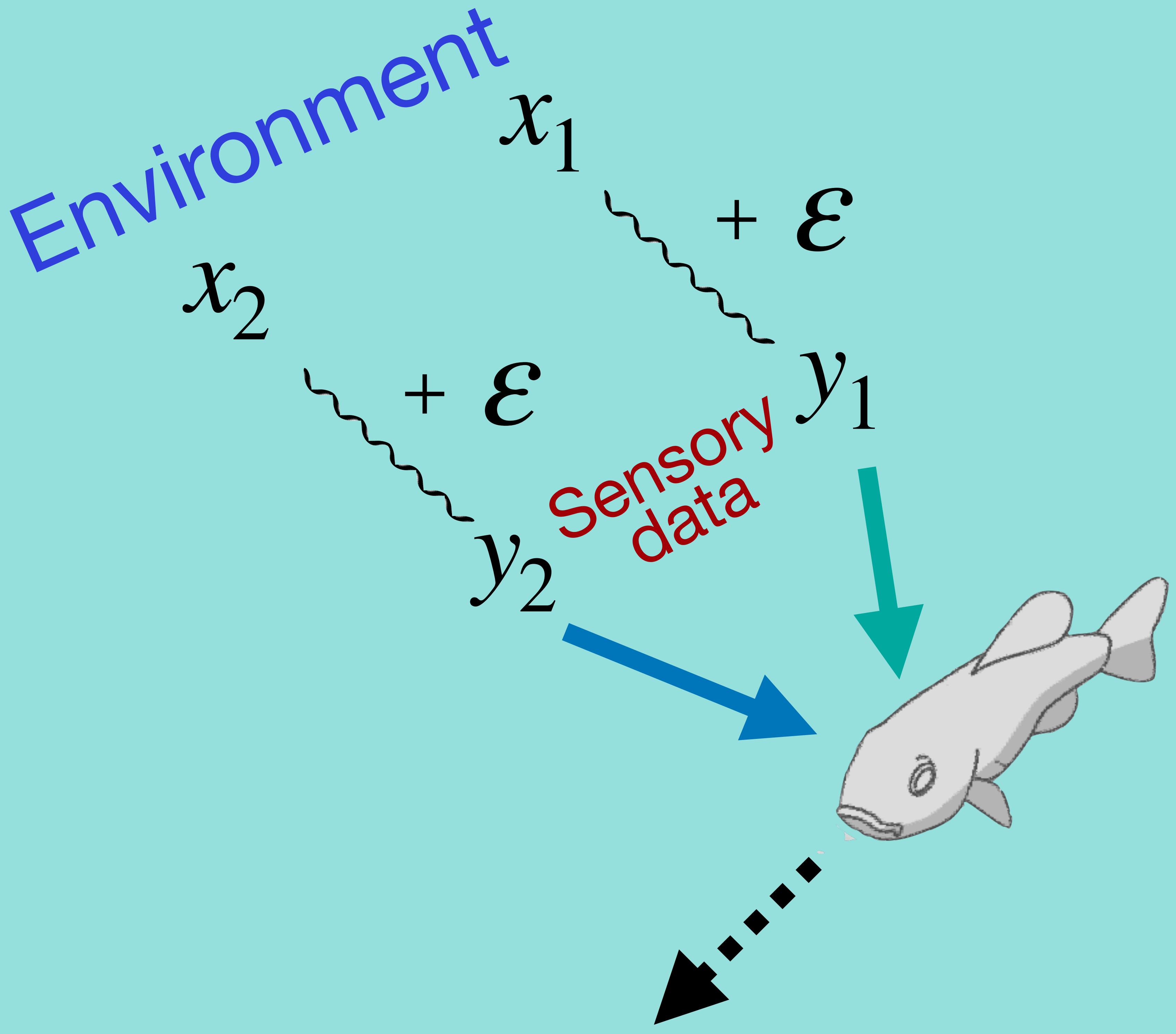


Comparing classical and Bayesian approaches

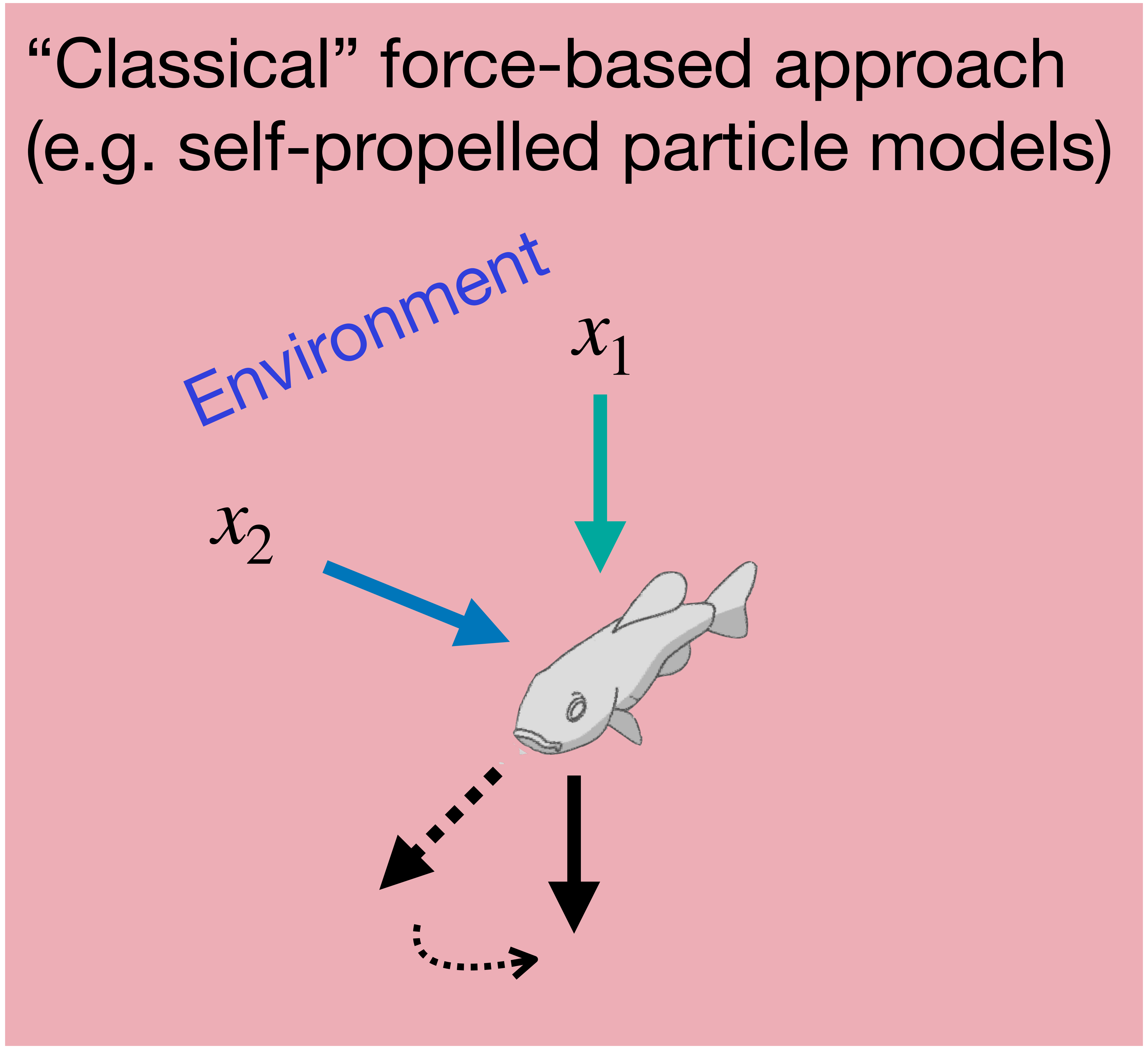
"Classical" force-based approach
(e.g. self-propelled particle models)

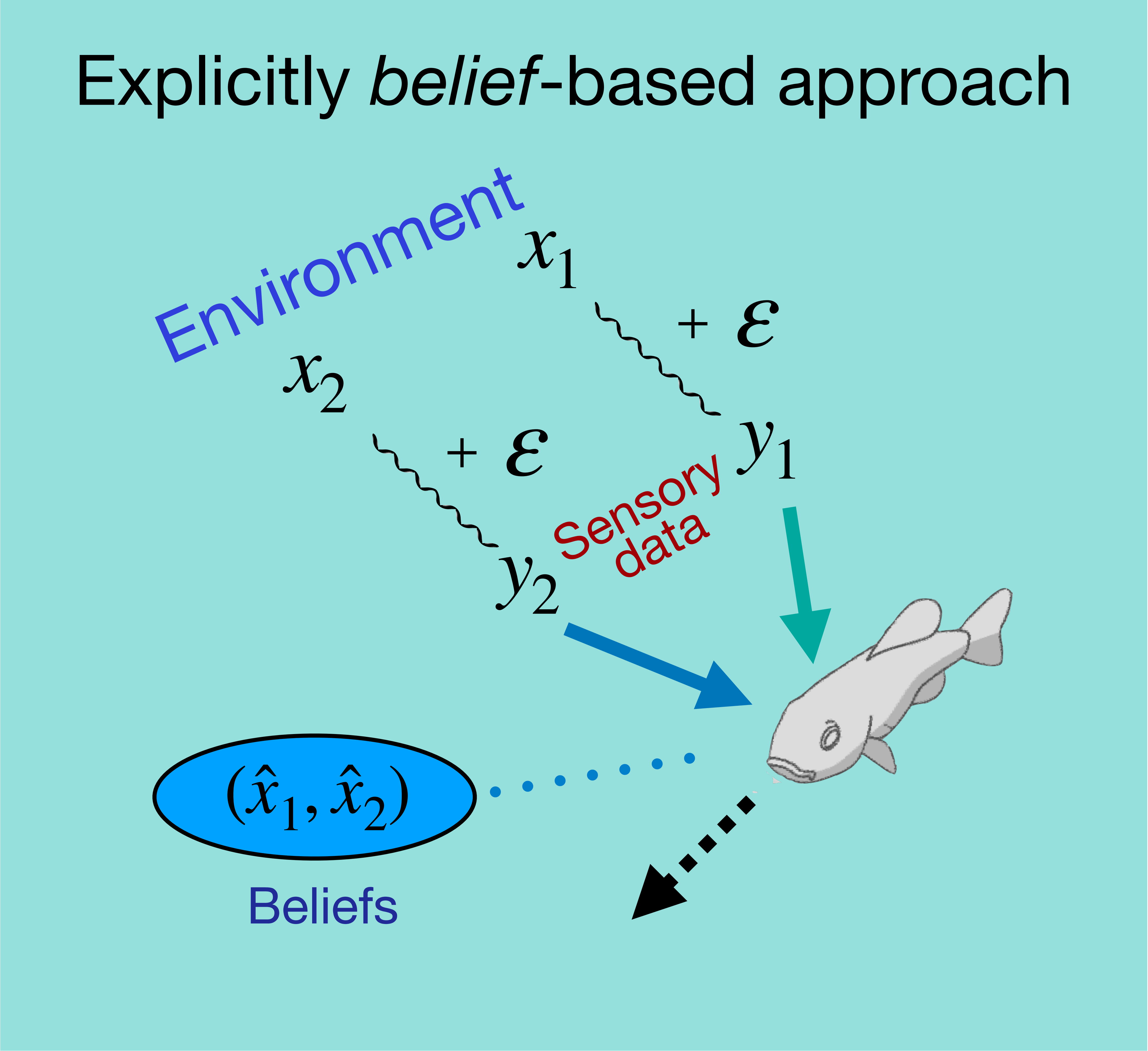


Explicitly *belief*-based approach



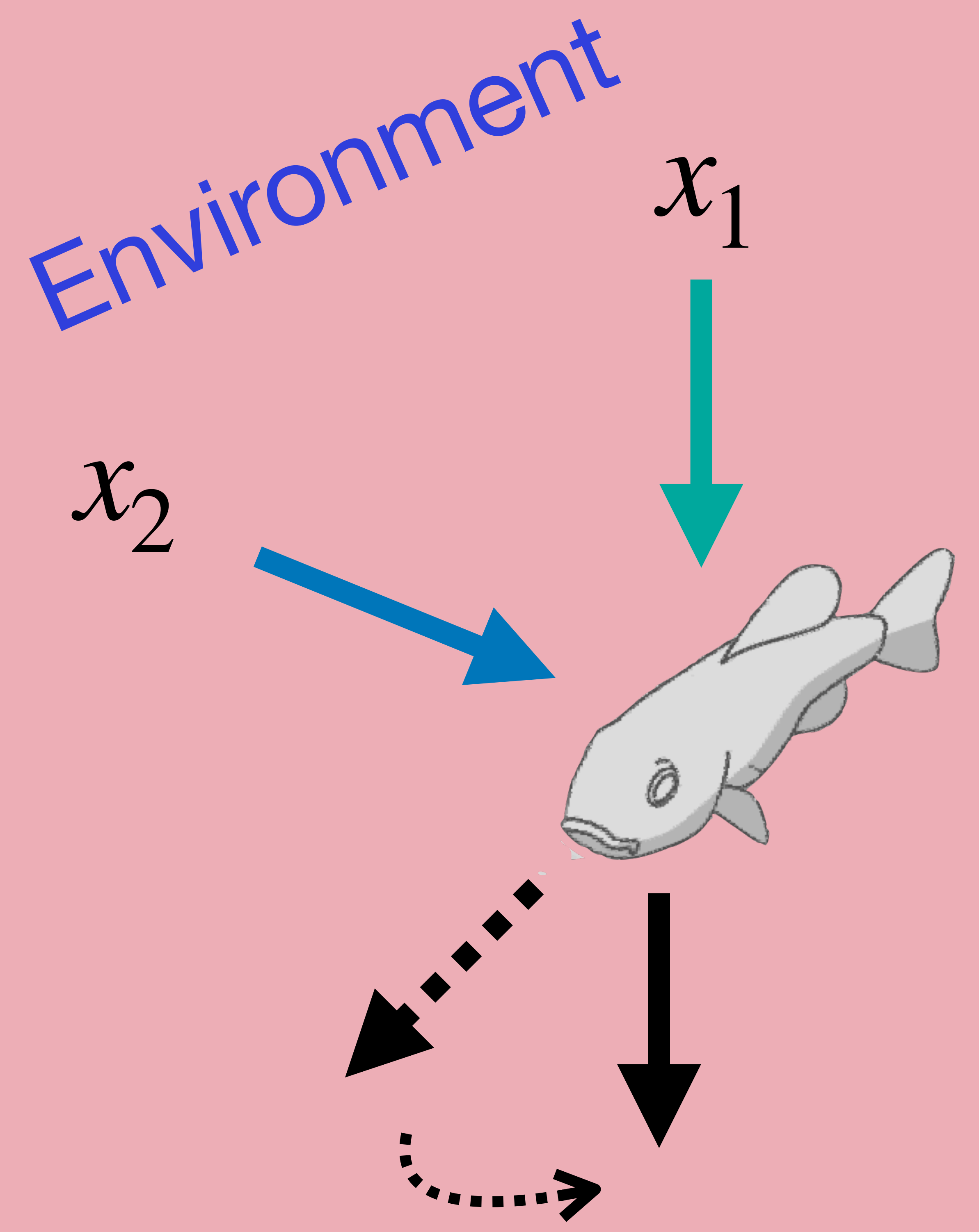
Comparing classical and Bayesian approaches



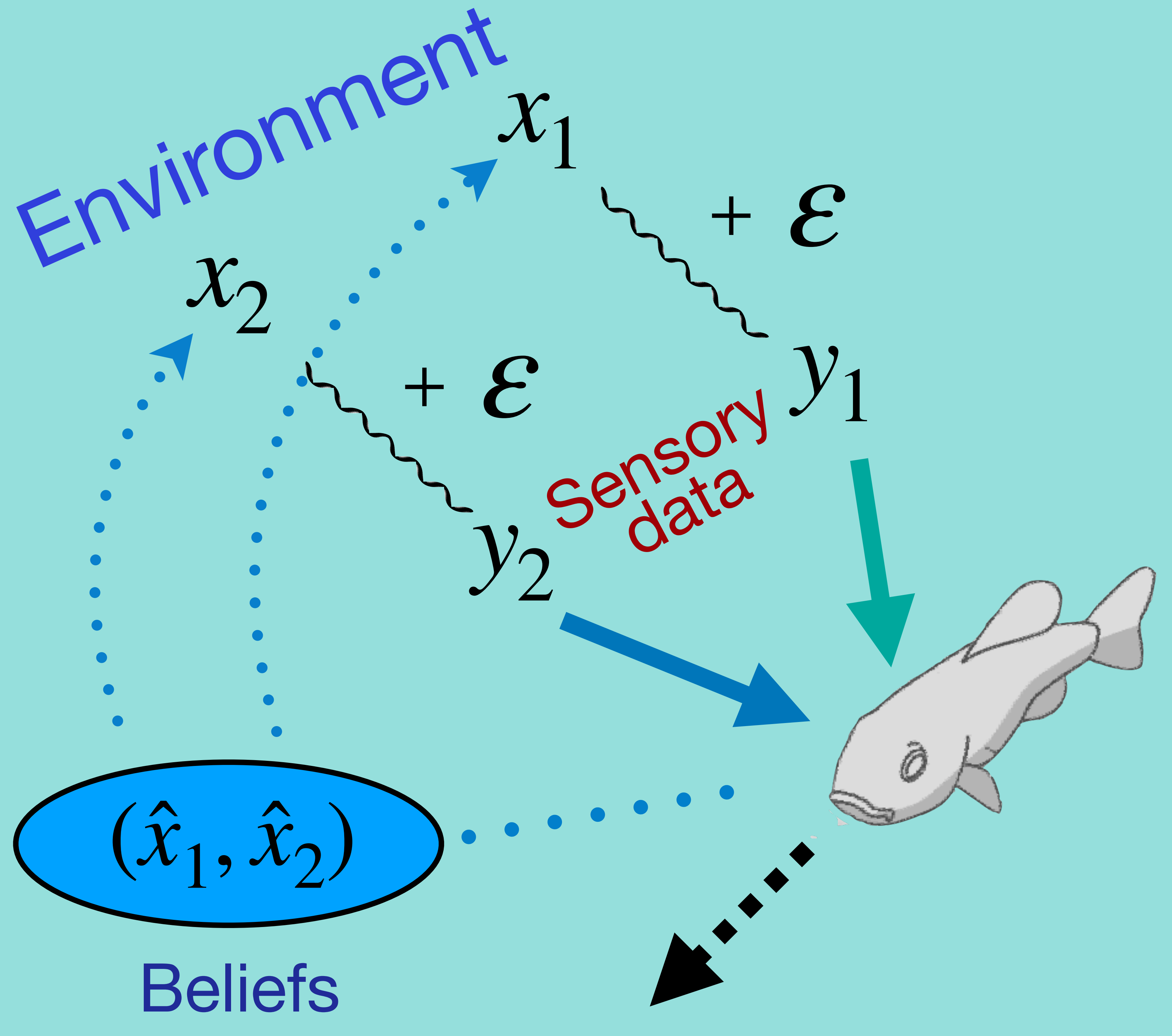


Comparing classical and Bayesian approaches

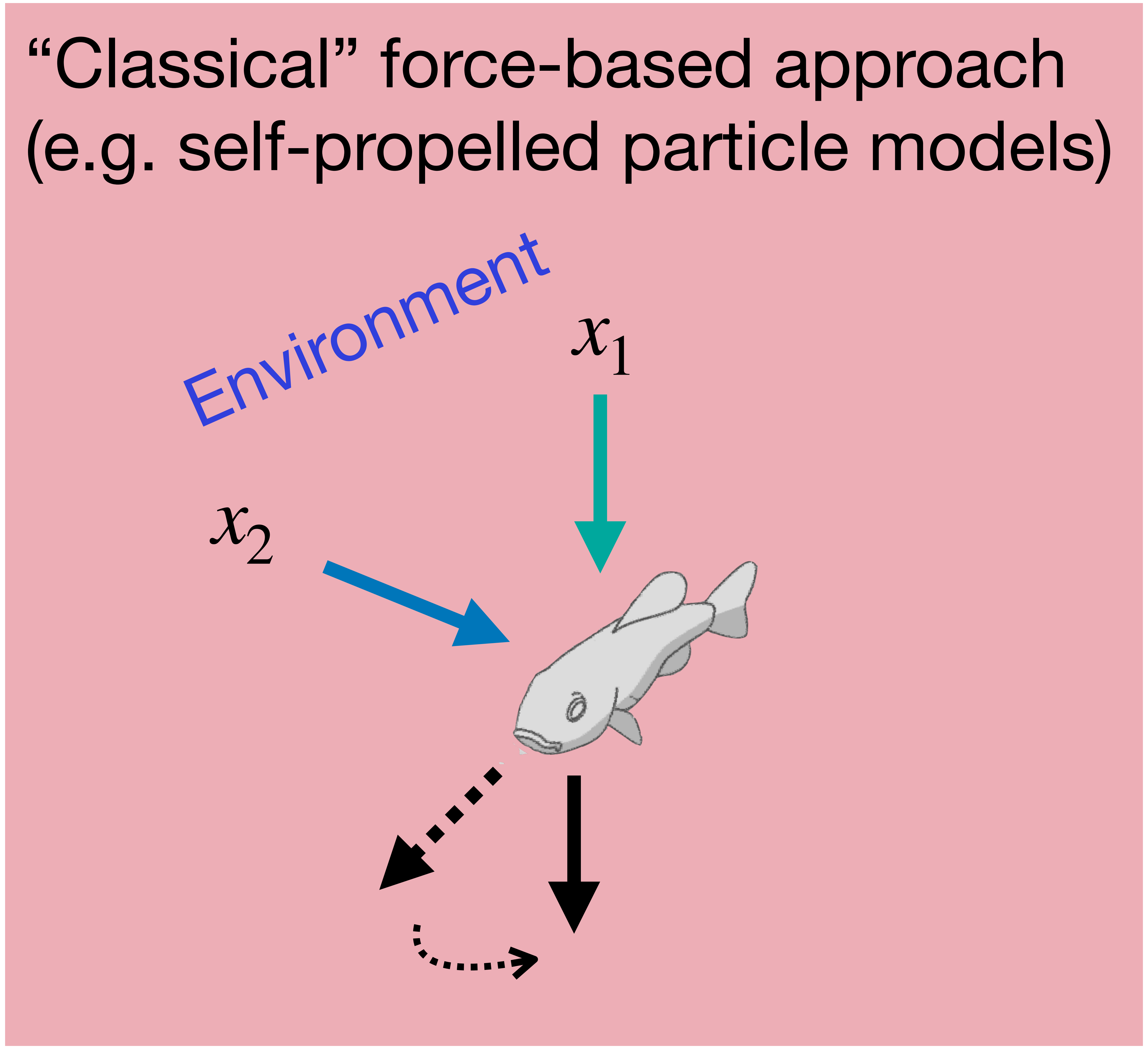
“Classical” force-based approach (e.g. self-propelled particle models)

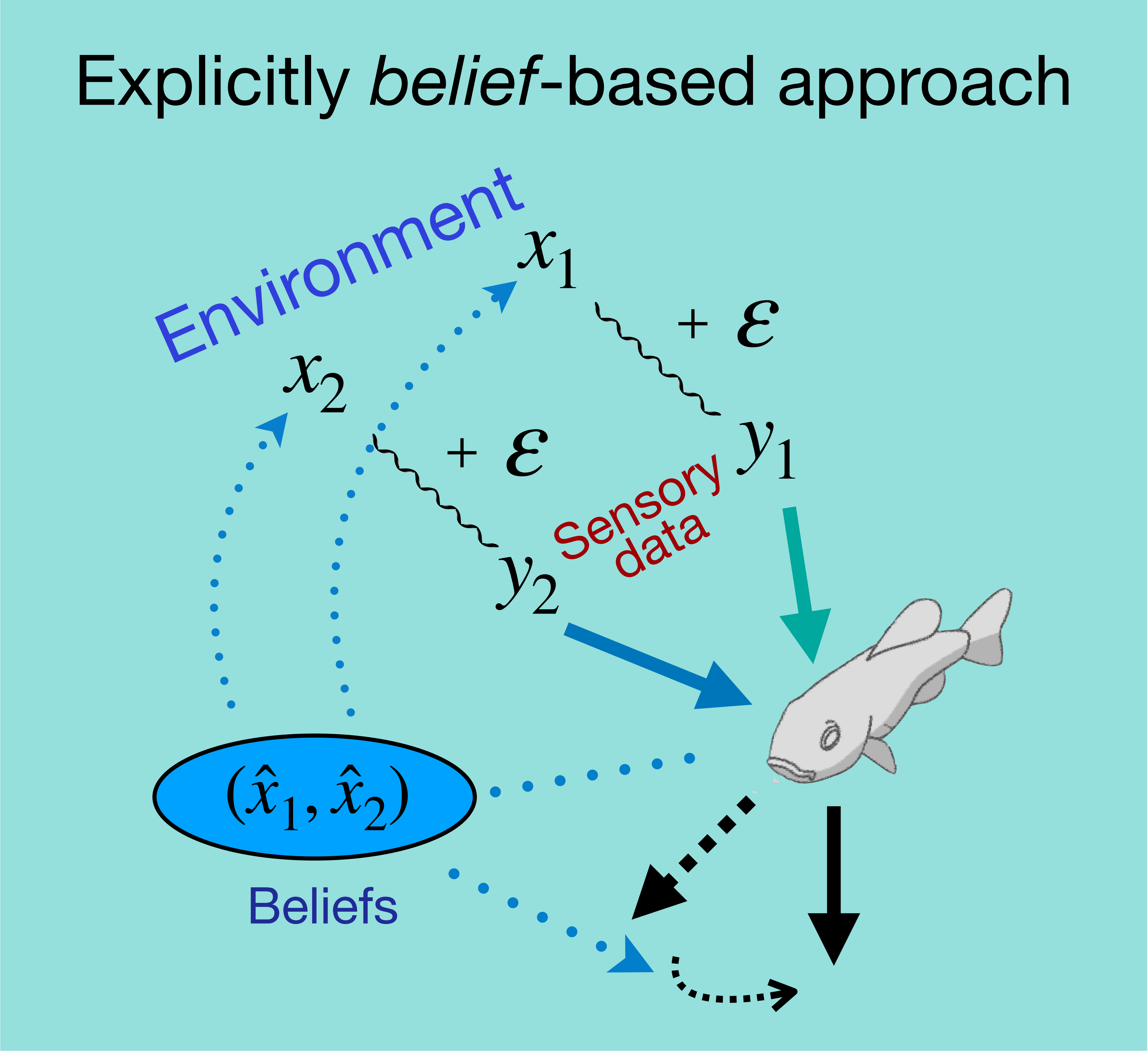


Explicitly *belief*-based approach



Comparing classical and Bayesian approaches





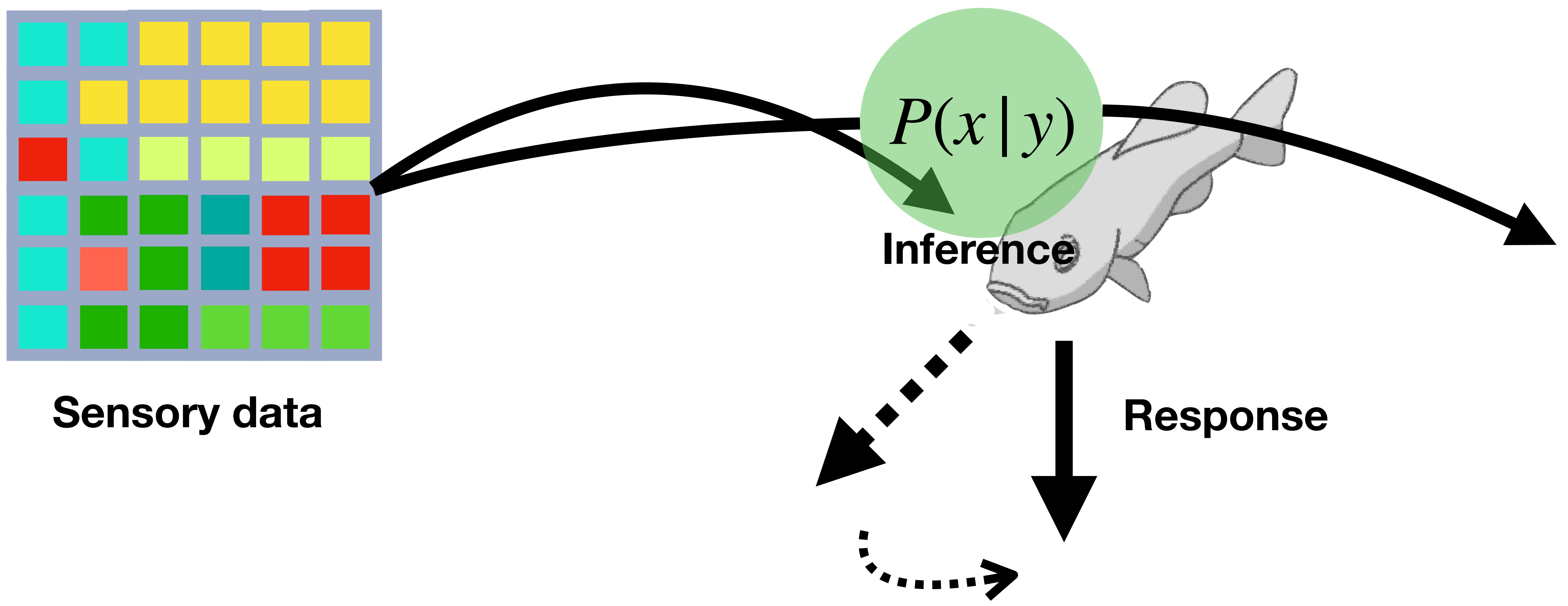
SCIENCE ADVANCES | RESEARCH ARTICLE

PHYSICS

A model of collective behavior based purely on vision

Renaud Bastien[1,2*†] and Pawel Romanczuk[3,4†]

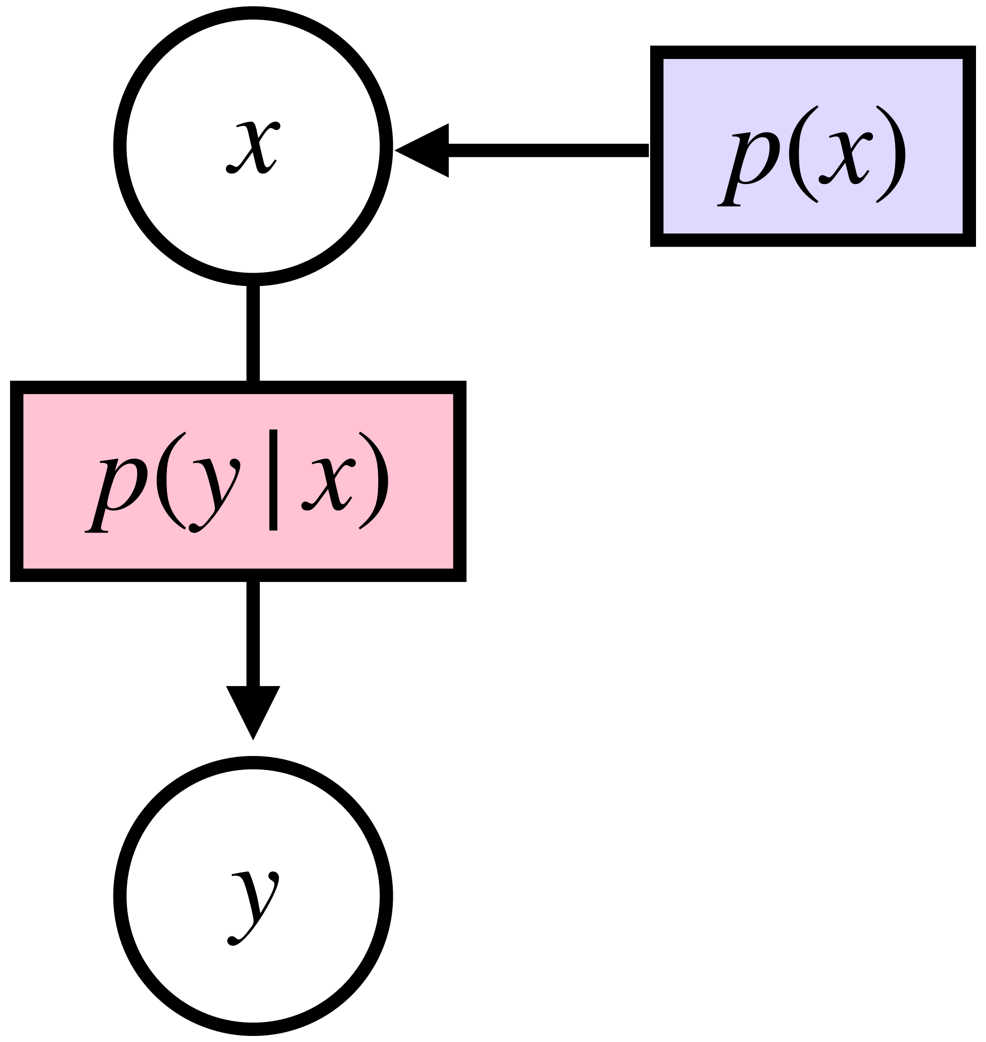
"Collective behavior crucially depends on the sensory information available to individuals; thus, ***ignoring perception by relying on ad hoc rules strongly limits our understanding of the underlying complexity of the problem****. Besides, it obstructs the interdisciplinary exchange between biology, neuroscience, engineering, and physics."*



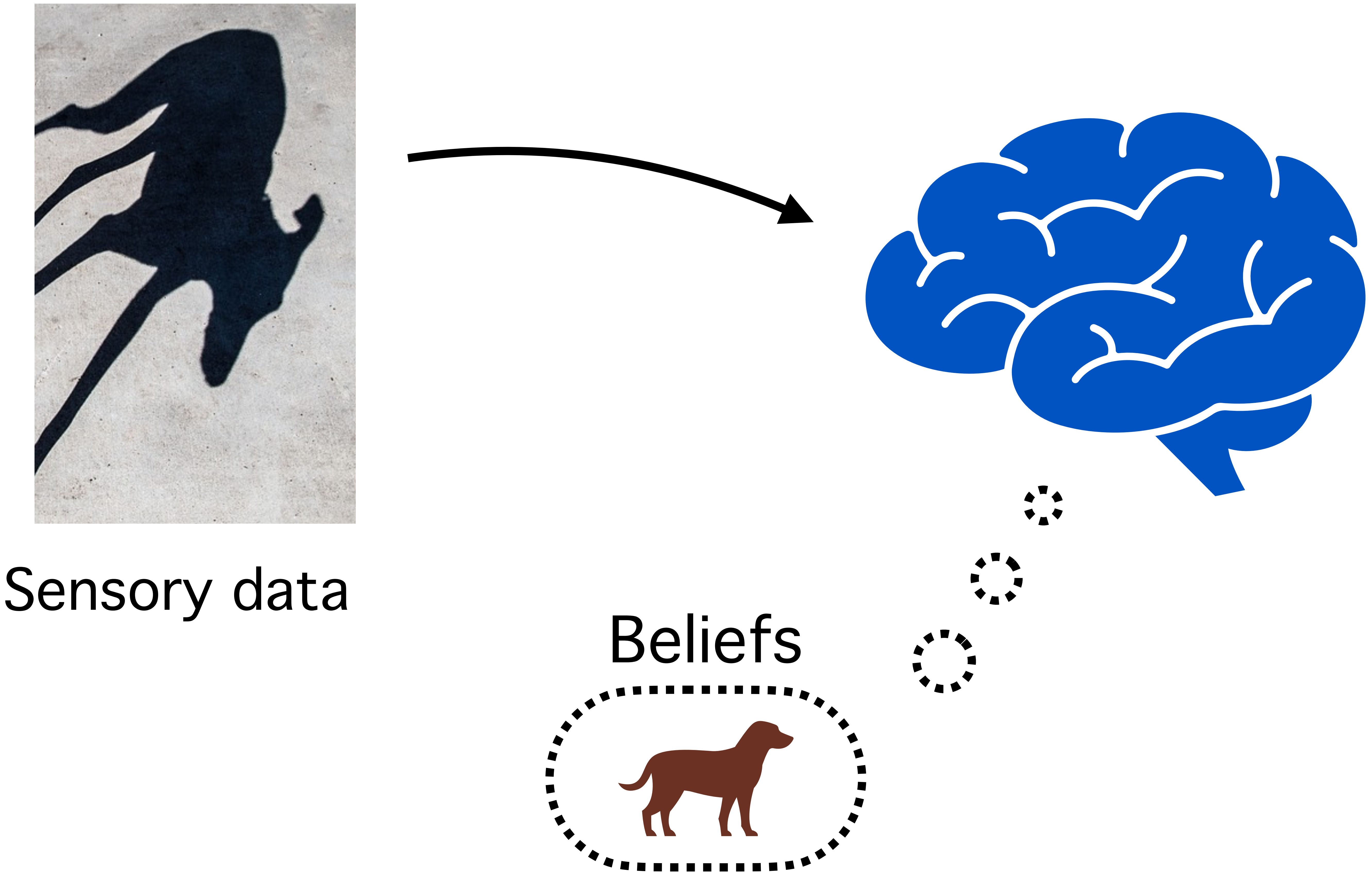
Writing down an agent's world model
a.k.a. the generative model

Generative model

$$p(y, x) = p(y \mid x)p(x)$$

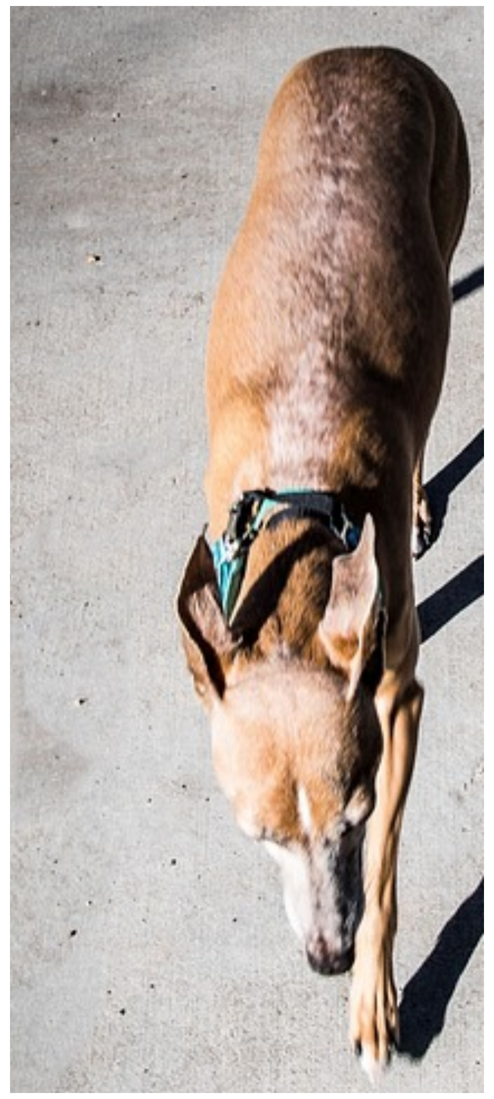


Generative model needed for inference



Generative model needed for inference

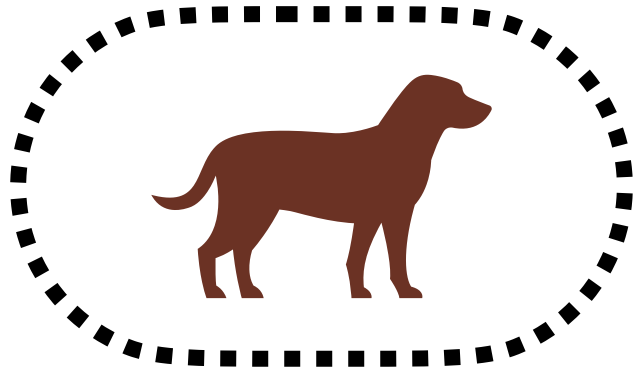
Generative model captures in-built assumptions about optics, light refraction, prevalence of objects, etc.



Hidden states

Sensory data

Beliefs



Writing down an agent's internal model a.k.a. the generative model

Generative model

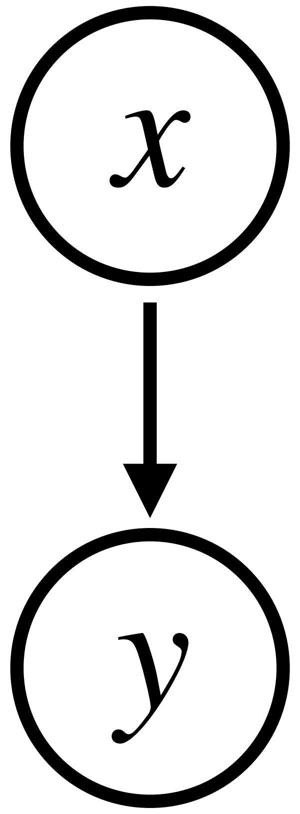
$$p(y, x) = p(y \mid x)p(x)$$

What sort of generative model might an individual in a mobile group have?

Generative model for an individual

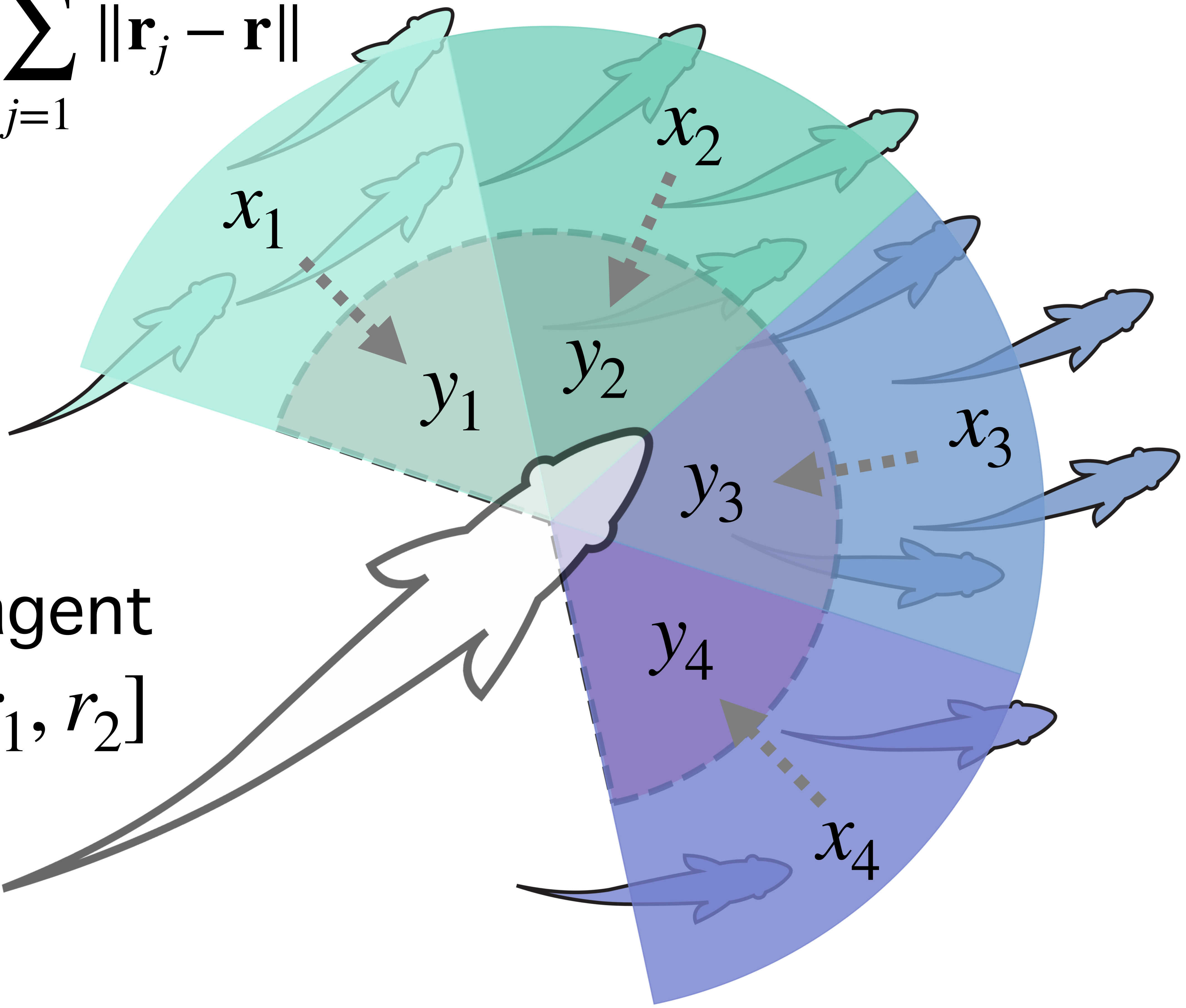
$$p(y, x) = p(y \mid x)p(x)$$

Hidden states x_t comprise the agent's environment



Sector-specific average distance

$$x_l = \frac{1}{K_l} \sum_{j=1}^{K_l} \|\mathbf{r}_j - \mathbf{r}\|$$



Focal agent

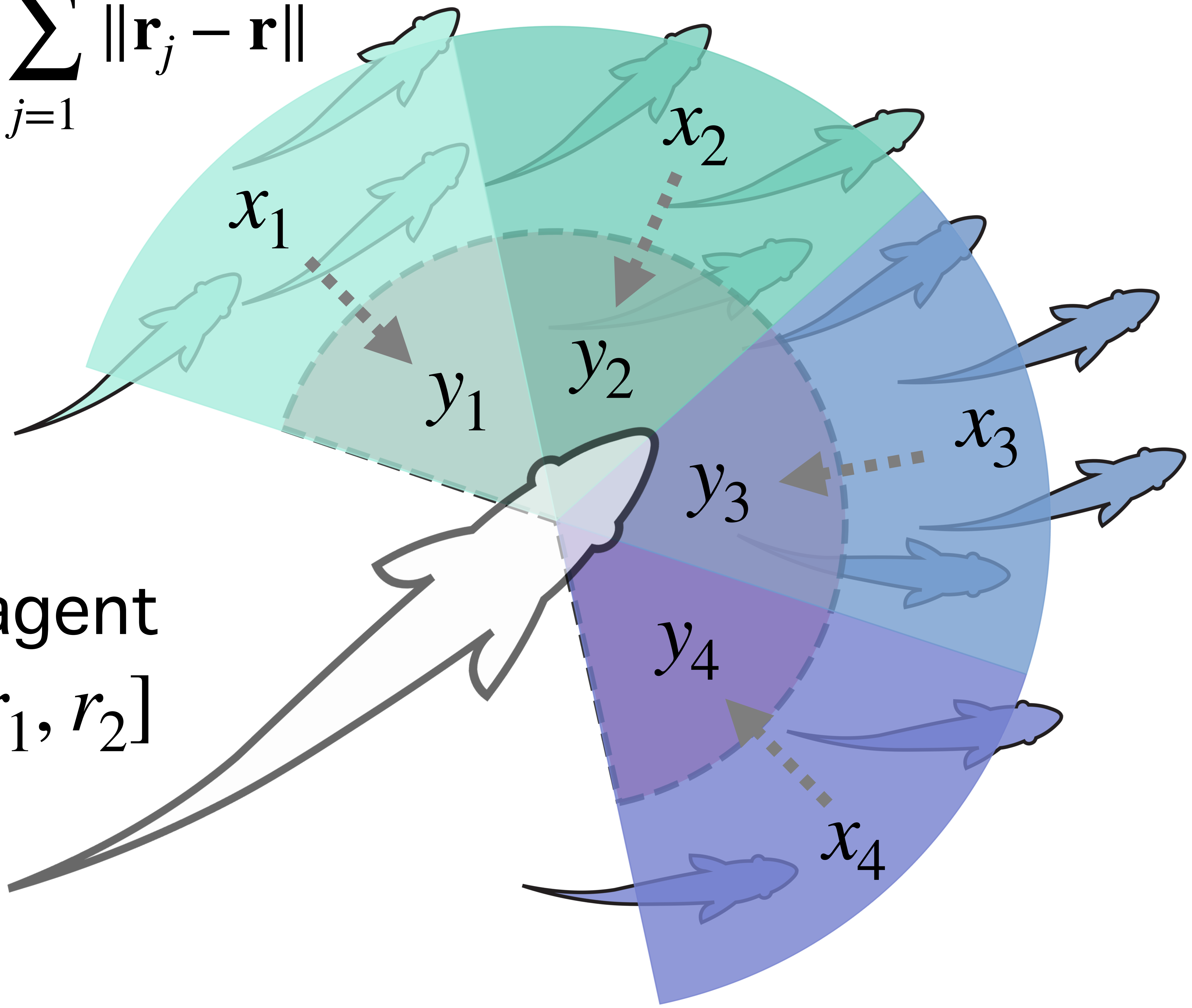
$\mathbf{r} = [r_1, r_2]$

Priors about social distance x_l

$$P(x_l) = N(\eta_l, \sigma_\omega)$$

Sector-specific average distance

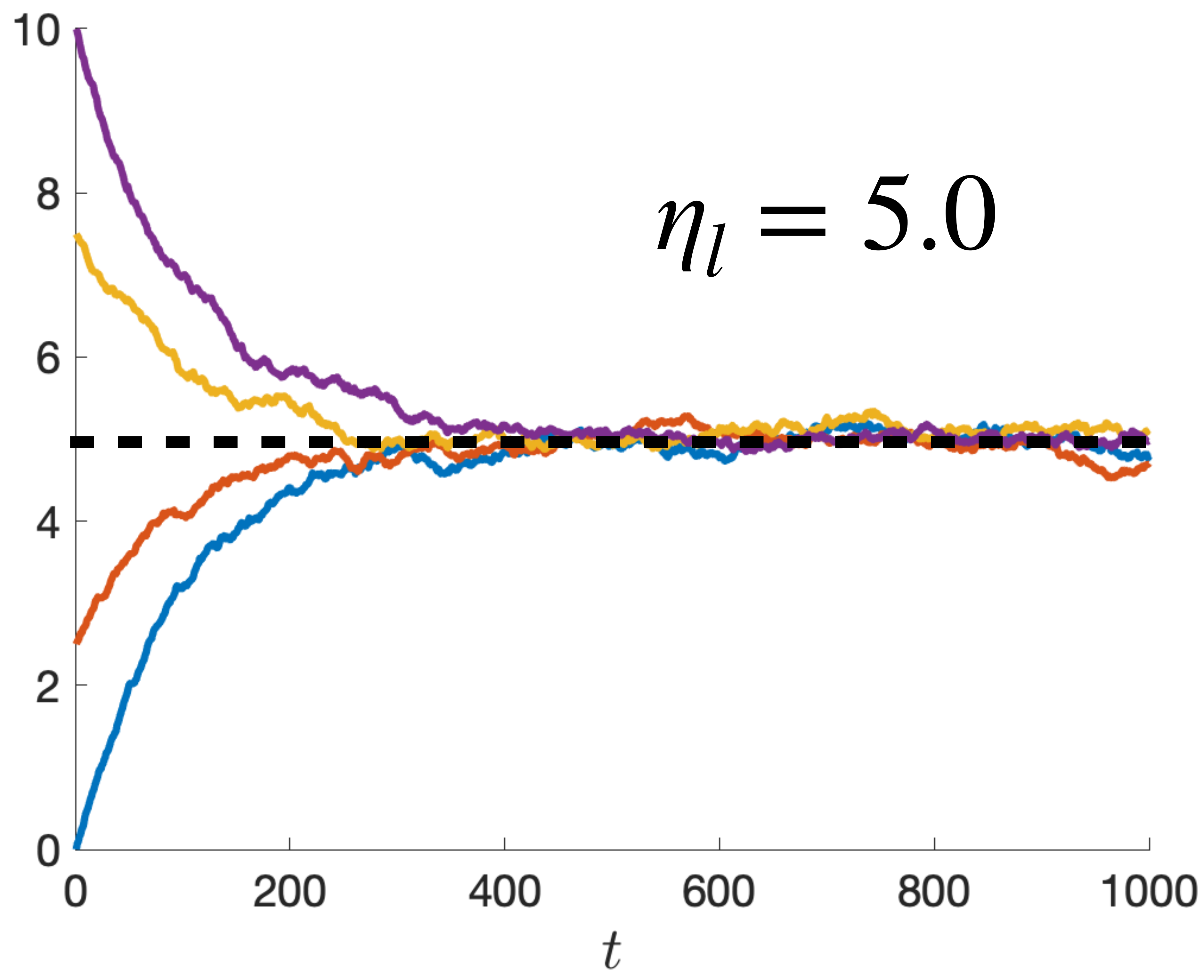
$$x_l = \frac{1}{K_l}\sum_{j=1}^{K_l} \|\mathbf{r}_j - \mathbf{r}\|$$



Focal agent

$\mathbf{r} = [r_1, r_2]$

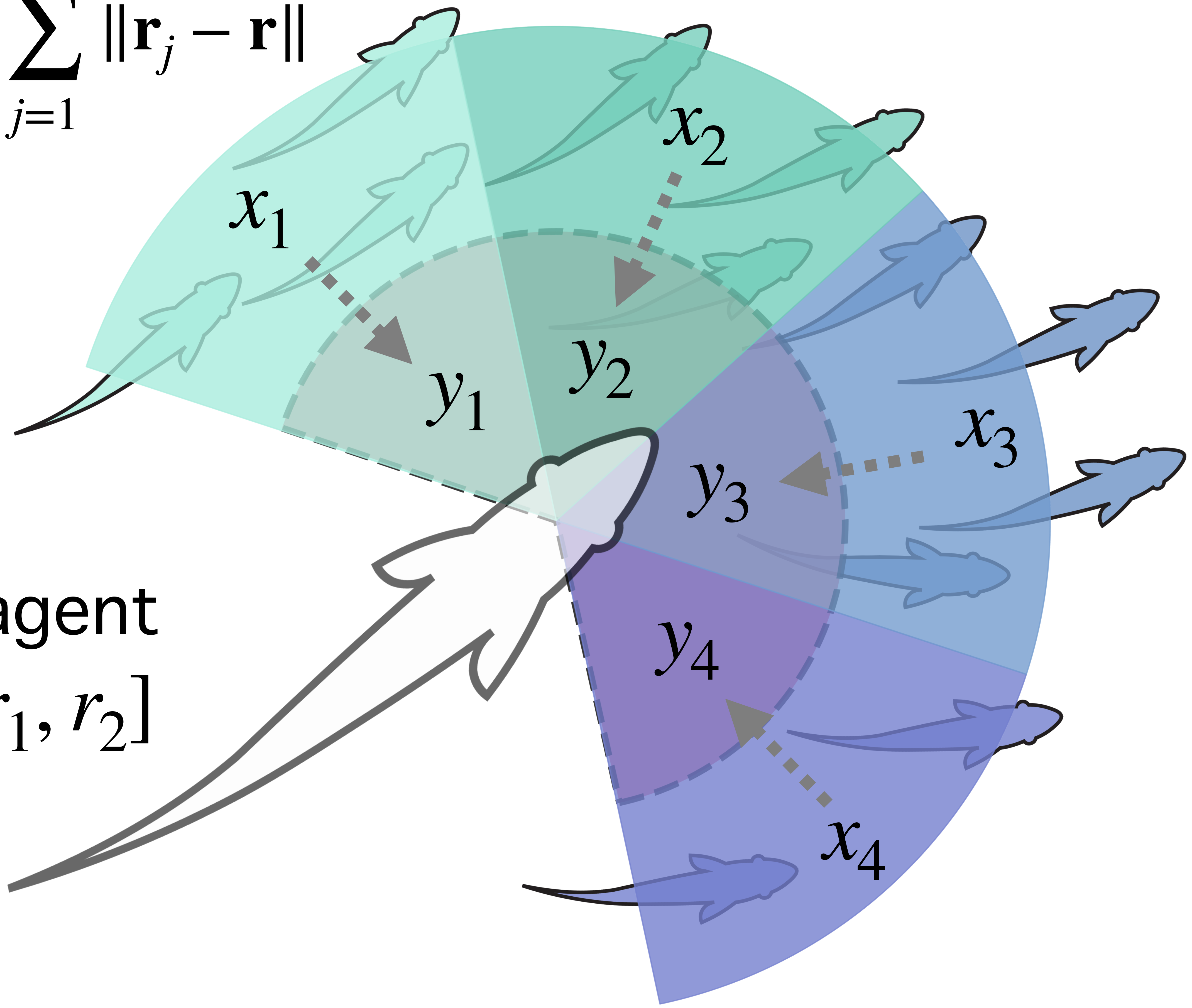
Prior belief about
the social distance in a particular sector



Social forces emerge from predictive control

Sector-specific average distance

$$x_l = \frac{1}{K_l}\sum_{j=1}^{K_l} \|\mathbf{r}_j - \mathbf{r}\|$$



Focal agent

$\mathbf{r} = [r_1, r_2]$

$\mathbf{v} =$ Direction vector

Sensorimotor contingency

Precision-weighted PE

$$\frac{d\mathbf{v}}{dt} = -\boxed{\frac{\partial y(\mathbf{v})}{\partial \mathbf{v}}}\boxed{\frac{\partial F}{\partial y(\mathbf{v})}}$$

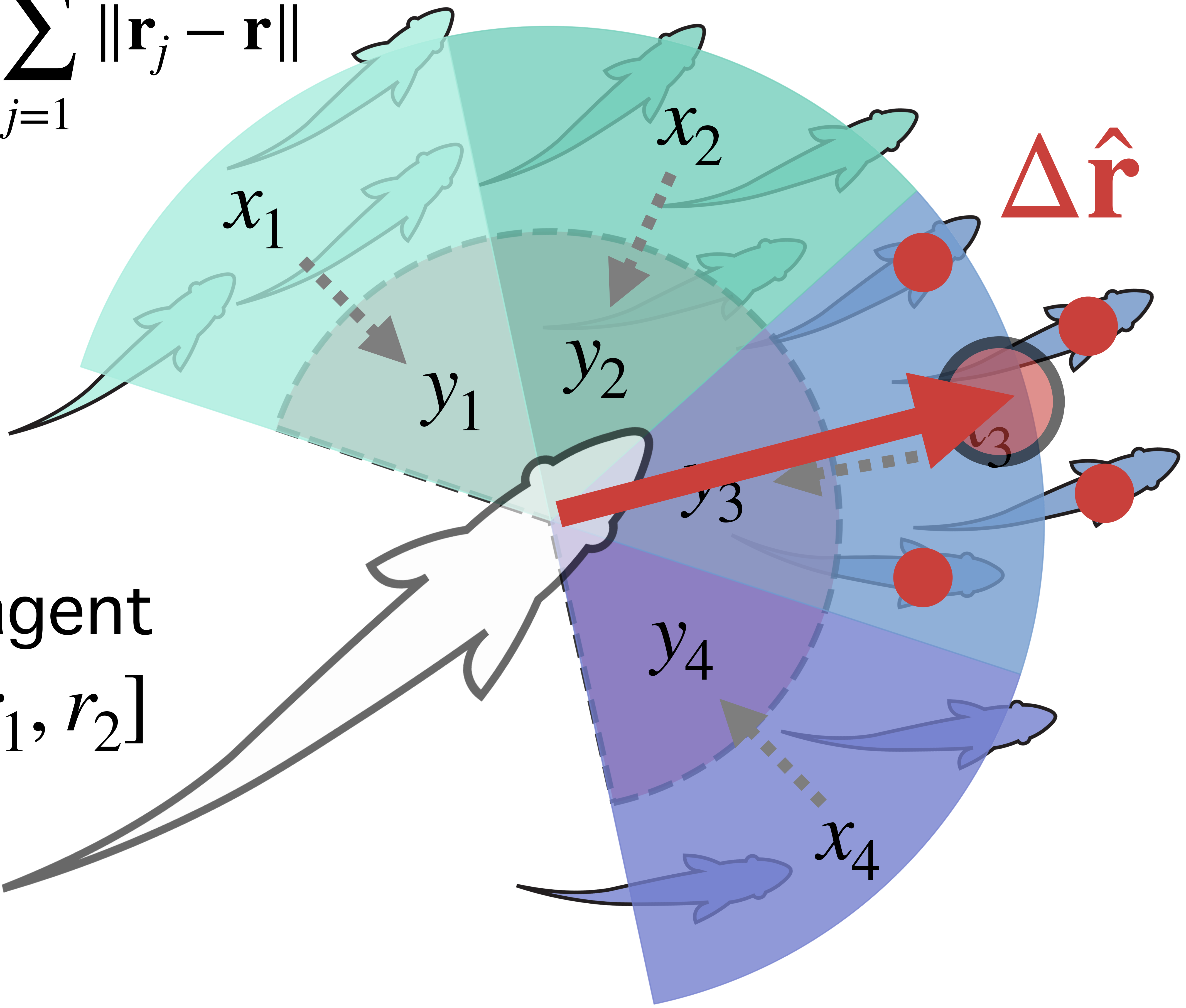
F doesn't directly depend on actions

(chain rule)

Social forces emerge from predictive control

Sector-specific average distance

$$x_l = \frac{1}{K_l}\sum_{j=1}^{K_l} \|\mathbf{r}_j - \mathbf{r}\|$$



Focal agent

$\mathbf{r} = [r_1, r_2]$

$\mathbf{v} =$ Direction vector

Sensorimotor contingency

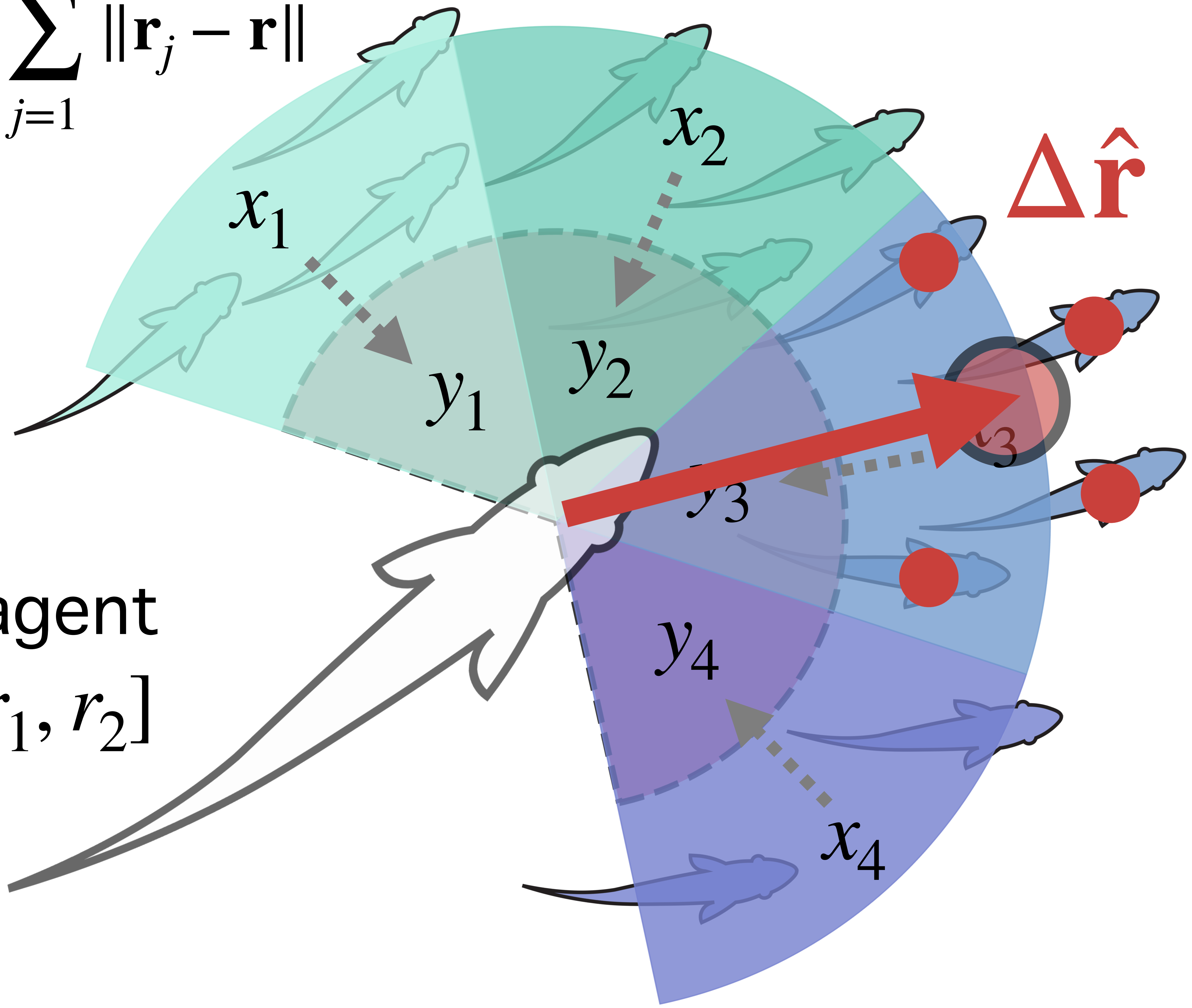
Precision-weighted PE

$$\frac{d\mathbf{v}}{dt} = \Delta\hat{\mathbf{r}}^{\top}\pi_z(y_l - \mu_l)$$

Social forces emerge from predictive control

Sector-specific average distance

$$x_l = \frac{1}{K_l} \sum_{j=1}^{K_l} \|\mathbf{r}_j - \mathbf{r}\|$$



Focal agent

$\mathbf{r} = [r_1, r_2]$

$\mathbf{v}$ = Direction vector

Sensorimotor contingency

Precision-weighted PE

$$\frac{d\mathbf{v}}{dt} = \boxed{\Delta\hat{\mathbf{r}}^\top}\boxed{\pi_z(y_l - \mu_l)}$$

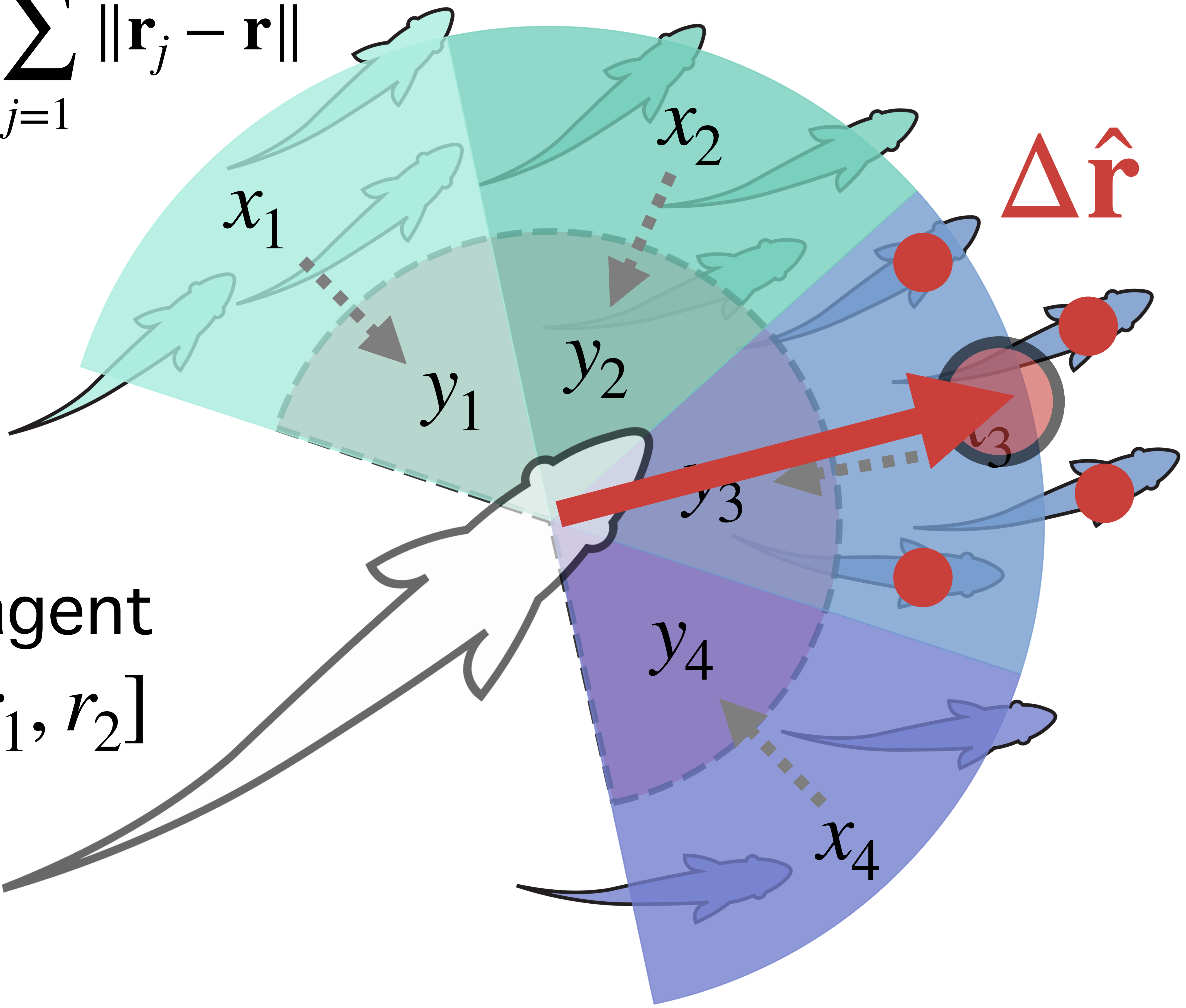
$y_l - \mu_l > 0$ ⟶ Attraction

$y_l - \mu_l < 0$ ⟶ Repulsion

Social forces emerge from predictive control

Sector-specific average distance

$$x_l = \frac{1}{K_l} \sum_{j=1}^{K_l} \|\mathbf{r}_j - \mathbf{r}\|$$



Focal agent

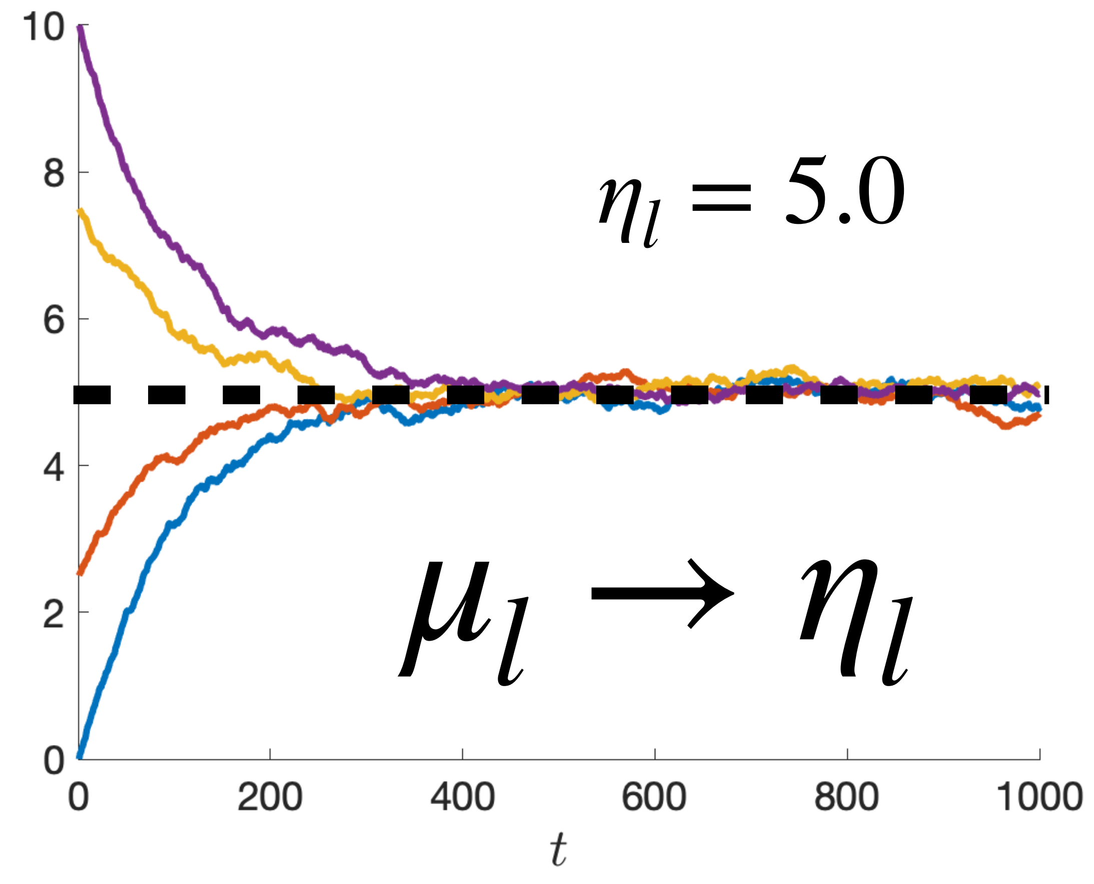
$\mathbf{r} = [r_1, r_2]$

$\mathbf{v}$ = Direction vector

Sensorimotor contingency

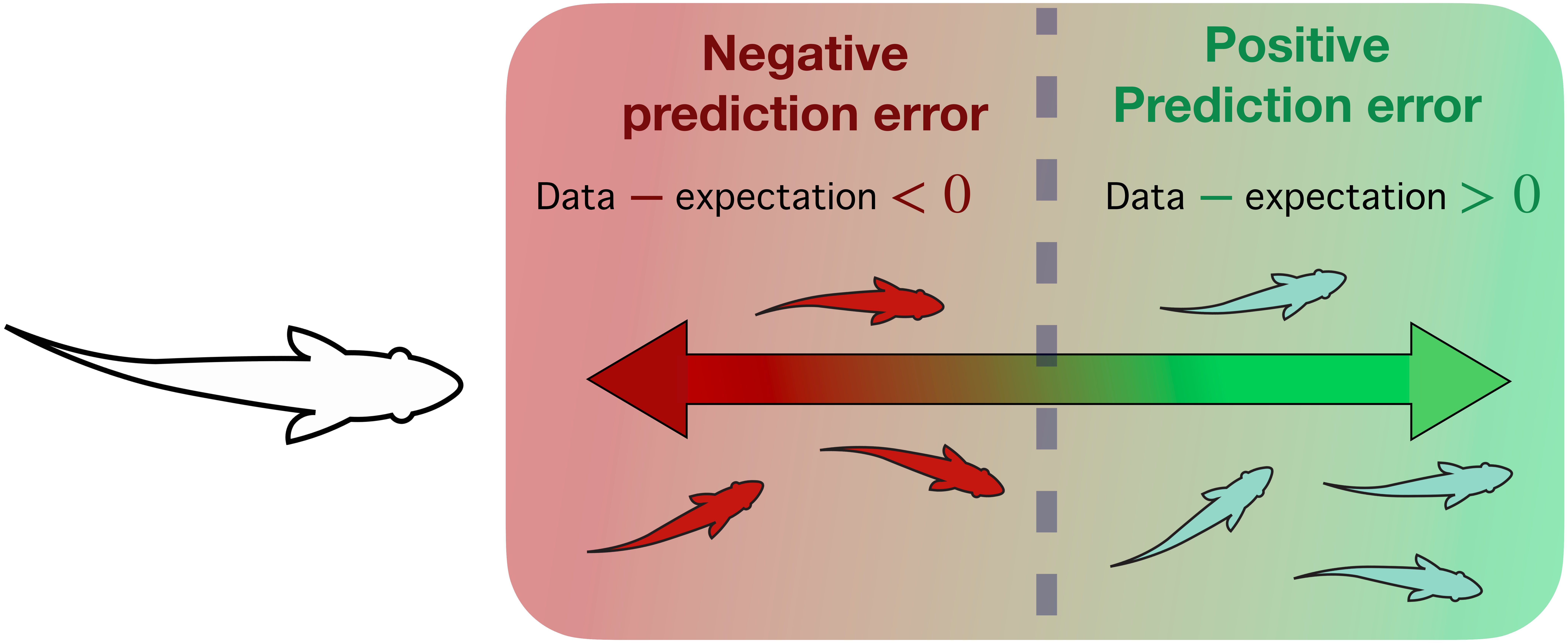
Precision-weighted PE

$$\frac{d\mathbf{v}}{dt} = \Delta\hat{\mathbf{r}}^\top \pi_z(y_l - \mu_l)$$

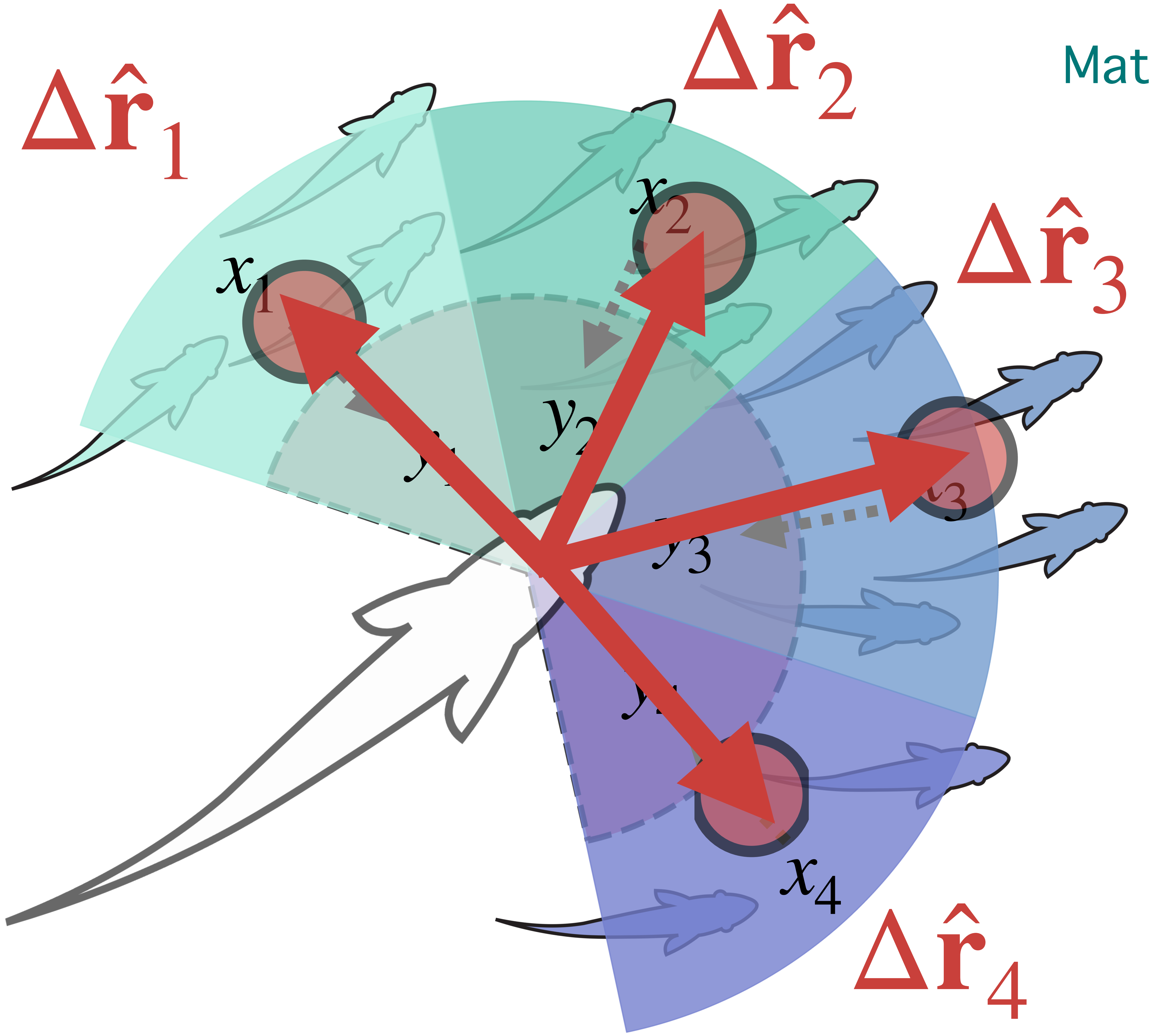


Zonal social force models

Active control of prediction error



Social forces emerge from (multivariate) predictive control



Matrix of sector vectors

Sector-wise PEs

$$\frac{d\mathbf{v}}{dt} = \underbrace{\Delta\hat{\mathbf{R}}^\top}_{\text{Matrix of sector vectors}}\underbrace{\tilde{\boldsymbol{\xi}}_{\mathbf{y}}}_{\text{Sector-wise PEs}}$$

$$\Delta\hat{\mathbf{R}} = \begin{bmatrix} \Delta\hat{\mathbf{r}}_1 \\ \Delta\hat{\mathbf{r}}_2 \\ \vdots \\ \Delta\hat{\mathbf{r}}_L \end{bmatrix}$$

$$\tilde{\boldsymbol{\xi}}_{\mathbf{y}} = \begin{bmatrix} \nabla_{\tilde{y}_1} F \\ \nabla_{\tilde{y}_2} F \\ \vdots \\ \nabla_{\tilde{y}_L} F \end{bmatrix}$$

Important addendum
for collective motion theorists!

$$\frac{d\mathbf{v}}{dt} = \overbrace{\Delta\hat{\mathbf{r}}^{\top}}^{\text{Sensorimotor contingency}} \overbrace{\pi_z(y_l - \mu_l)}^{\text{Precision-weighted PE}}$$

$$y_l - \mu_l > 0 \dashrightarrow \text{Attraction}$$

$$y_l - \mu_l < 0 \dashrightarrow \text{Repulsion}$$

$$\tilde{y} = \begin{bmatrix} y \\ \\ \\ \end{bmatrix} = \begin{array}{cl} \begin{bmatrix} y \\ \\ \\ \end{bmatrix} & \begin{array}{l} \text{Sensed distance} \\ \\ \\ \end{array} \end{array}$$

$$\frac{d\mathbf{v}}{dt} = \nabla_{\tilde{y}}\mathbf{v}^{\top}\tilde{\Pi}_z(\tilde{y}_l - \tilde{\mu}_l)$$

$$y_l' - \mu_l' > 0 \dashrightarrow \text{Attraction}$$

$$y_l' - \mu_l' < 0 \dashrightarrow \text{Repulsion}$$

Important addendum
for collective motion theorists!

$$\tilde{y} = \begin{bmatrix} y \\ y' \\ y'' \\ \vdots \end{bmatrix} = \begin{bmatrix} y \\ \partial_t y \\ \partial_t^2 y \\ \vdots \end{bmatrix} \begin{array}{l} \text{Sensed distance} \\ \text{Sensed “distance velocity”} \\ \text{Sensed “distance acceleration”} \\ \\ \end{array}$$

$$\frac{d\mathbf{v}}{dt} = \nabla_{\tilde{y}} \mathbf{v}^\top \tilde{\Pi}_z (\tilde{y}_l - \tilde{\mu}_l)$$

$$y_l' - \mu_l' > 0 \dashrightarrow \text{Attraction}$$

$$y_l' - \mu_l' < 0 \dashrightarrow \text{Repulsion}$$

Research article

Swarming and pattern formation due to selective attraction and repulsion

Pawel Romanczuk and Lutz Schimansky-Geier

Published: 26 September 2012 https://doi.org/10.1098/rsfs.2012.0030

$y_l' = \dfrac{dy_l'}{dt}$ is equivalent to the "relative velocity", or the rate at which neighbouring individuals are receding (positive) vs. looming (negative)

$$\frac{dy_l}{dt} = (\mathbf{r}_i - \mathbf{r}_j) \cdot \mathbf{v}_i + \sum_{j \in S_l} \Big((\mathbf{r}_j - \mathbf{r}_i) \cdot \mathbf{v}_j \Big)$$

$\mathbf{r}_i =$ Position vector of focal agent i

$\mathbf{r}_j =$ Position vector of neighbour j in sector l

Evidence for use of prediction errors (i.e., unpredicted changes in sensory input), rather than absolute values

nature communications

Article

https://doi.org/10.1038/s41467-024-53361-8

Body orientation change of neighbors leads to scale-free correlation in collective motion

Zhicheng Zheng[1], Yuan Tao[1], Yalun Xiang[1], Xiaokang Lei[2] & Xingguang Peng[1]

[1]School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, P. R. China. [2]School of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an, Shaanxi 710055, P. R. China. e-mail: pxg@nwpu.edu.cn

nature communications

Zebrafish capable of generating future state prediction error show improved active avoidance behavior in virtual reality

Makio Torigoe[1], Tanvir Islam[1,2], Hisaya Kakinuma[1,2], Chi Chung Alan Fung[3], Takuya Isomura[4], Hideaki Shimazaki[5], Tazu Aoki[1], Tomoki Fukai[3] & Hitoshi Okamoto[1,2]

PNAS RESEARCH ARTICLE | ENGINEERING BIOLOGICAL SCIENCES OPEN ACCESS

Individual error correction drives responsive self-assembly of army ant scaffolds

Matthew J. Lutz[a,b,c,2,1], Chris R. Reid[d,2,1], Christopher J. Lustri[e], Albert B. Kao[f], Simon Garnier[g], and Iain D. Couzin[a,b,c]

SCIENCE ADVANCES | RESEARCH ARTICLE

NEUROSCIENCE

Predictive neural computations in the cerebellum contribute to motor planning and faster behavioral responses in larval zebrafish

Sriram Narayanan, Aalok Varma, Vatsala Thirumalai*

Collective simulation achieved by minimizing individual free energy functionals

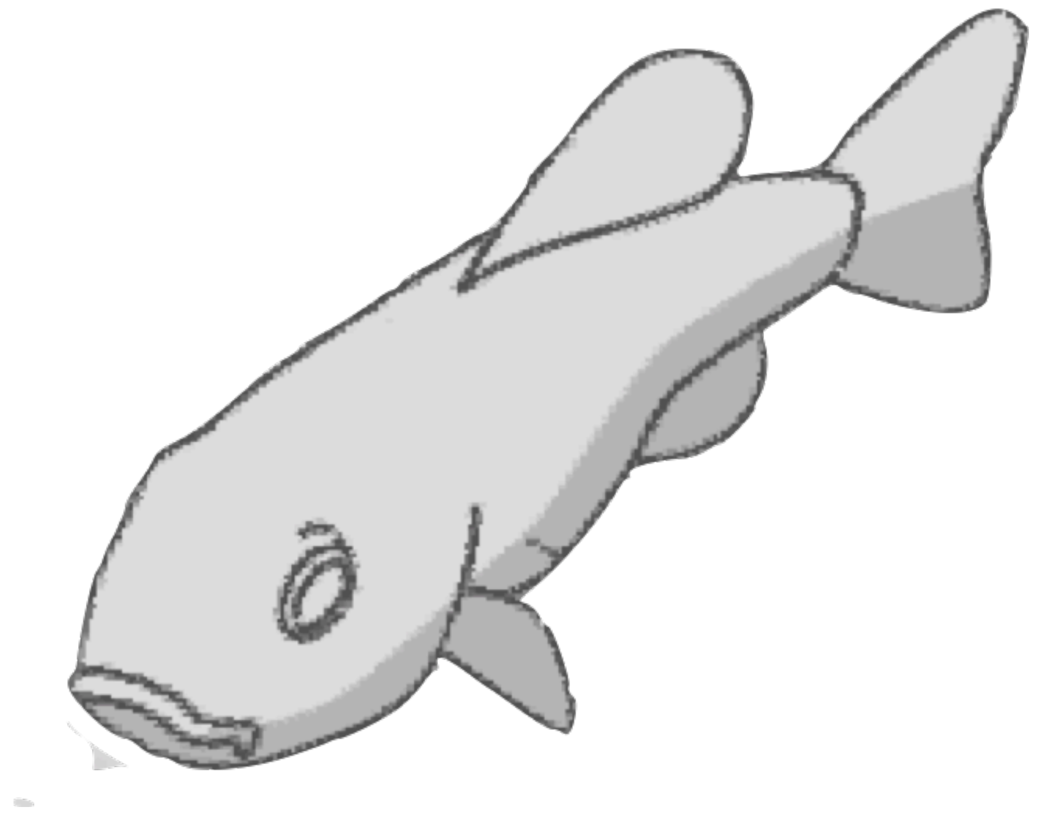
$$F = \text{Surprise} = \text{Prediction error}$$

$$\dot{\mu} = -\nabla_{\mu} F(\mu, y)$$

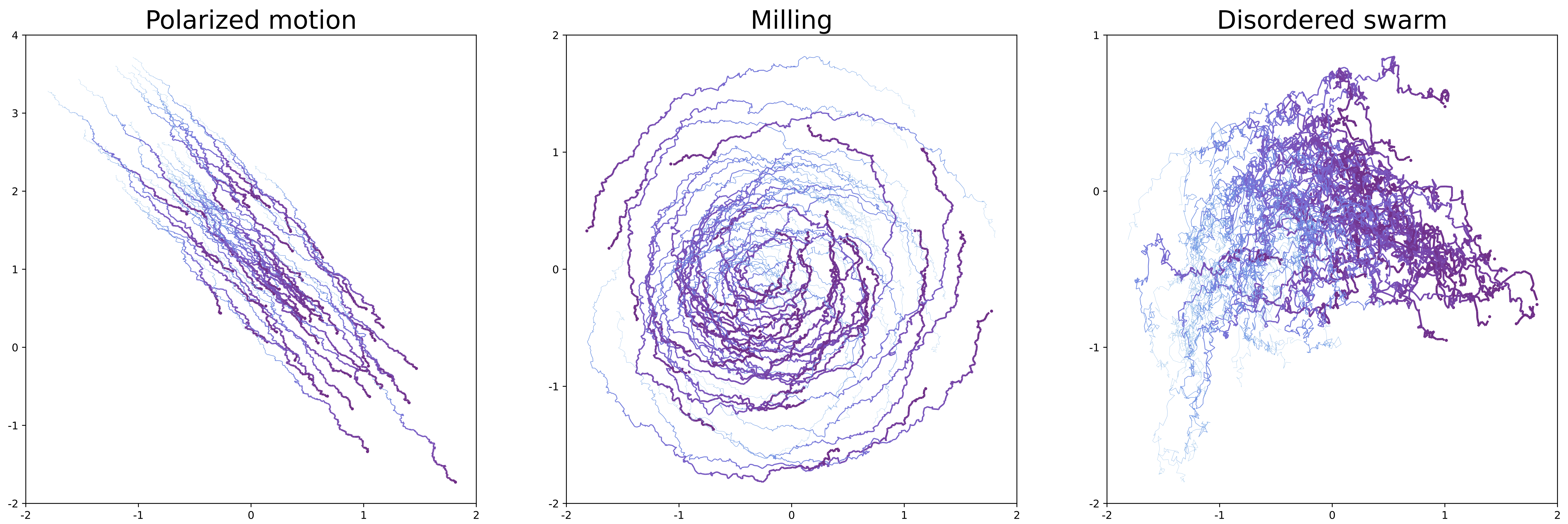
Perception

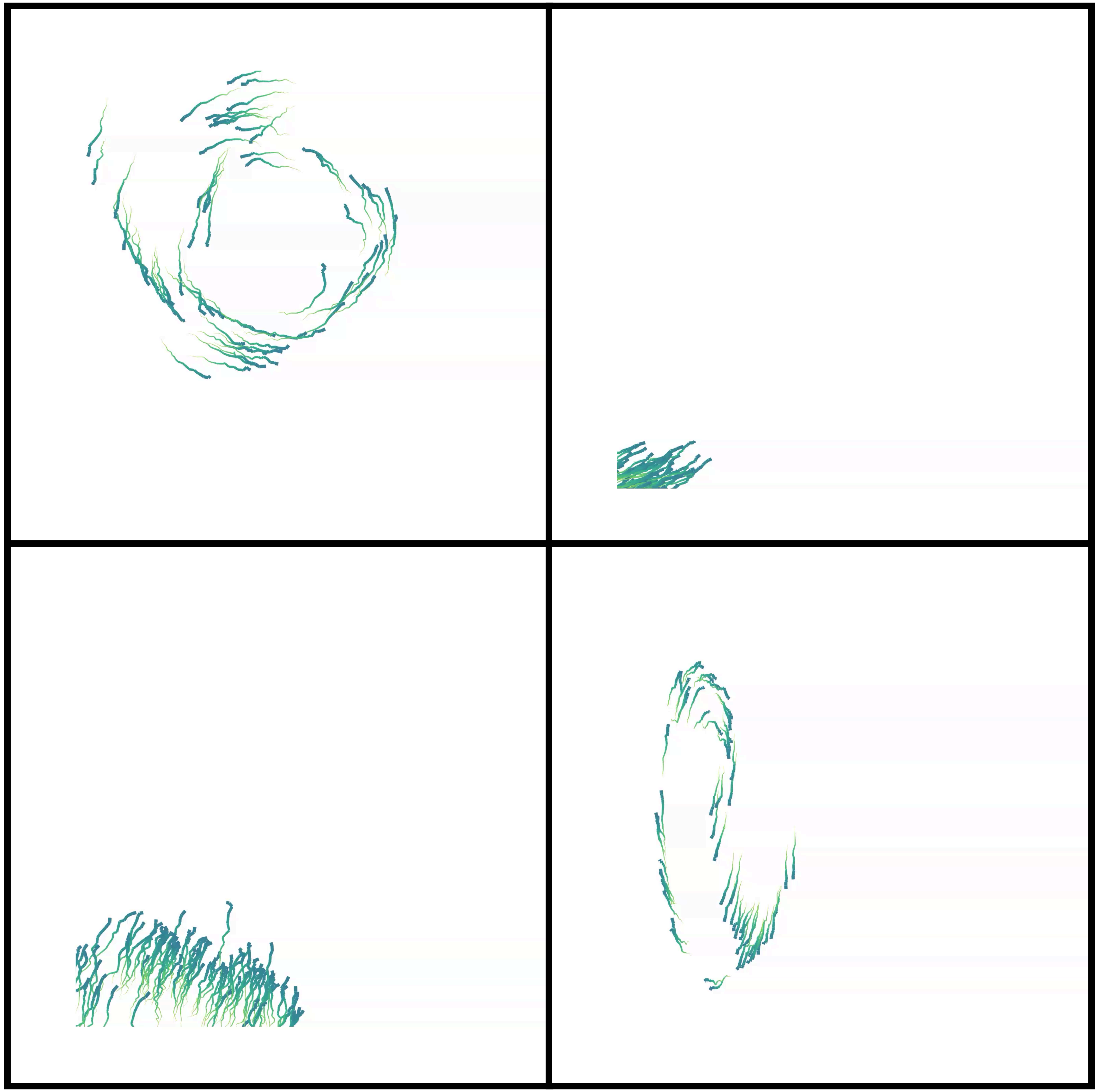
$$\dot{v} = -\nabla_{\mathbf{v}} F(\mu, y(\mathbf{v}))$$

Action



Collective regimes



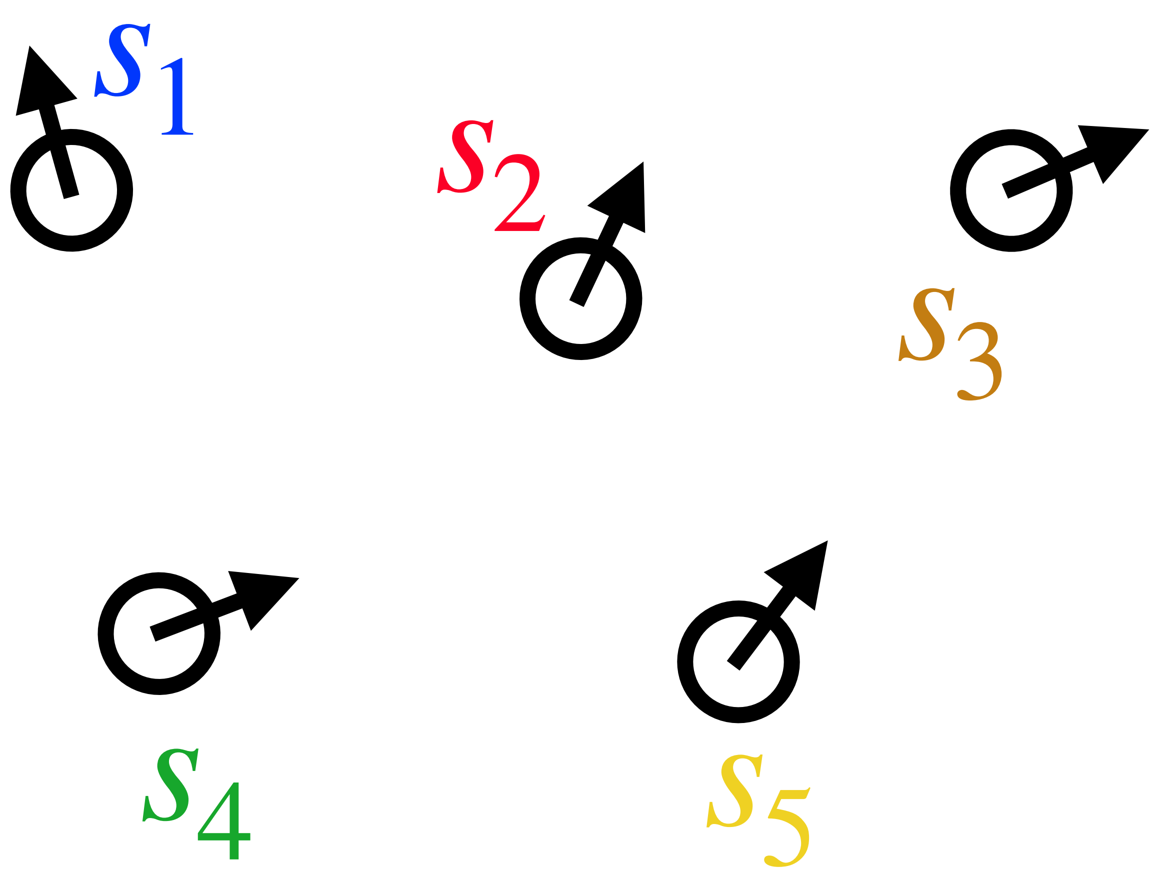


Potential function vs. free energy function(al)

Potential systems

- **Global function** of the configurational states of the system
- Individual dynamics move down gradients of the shared, global potential

$$E(s_1, s_2, s_3, \ldots)$$



$$\dot{s}_1 \propto -\nabla_{s_1} E$$

$$\dot{s}_2 \propto -\nabla_{s_2} E$$

$$\dot{s}_3 \propto -\nabla_{s_3} E$$

$$\vdots$$

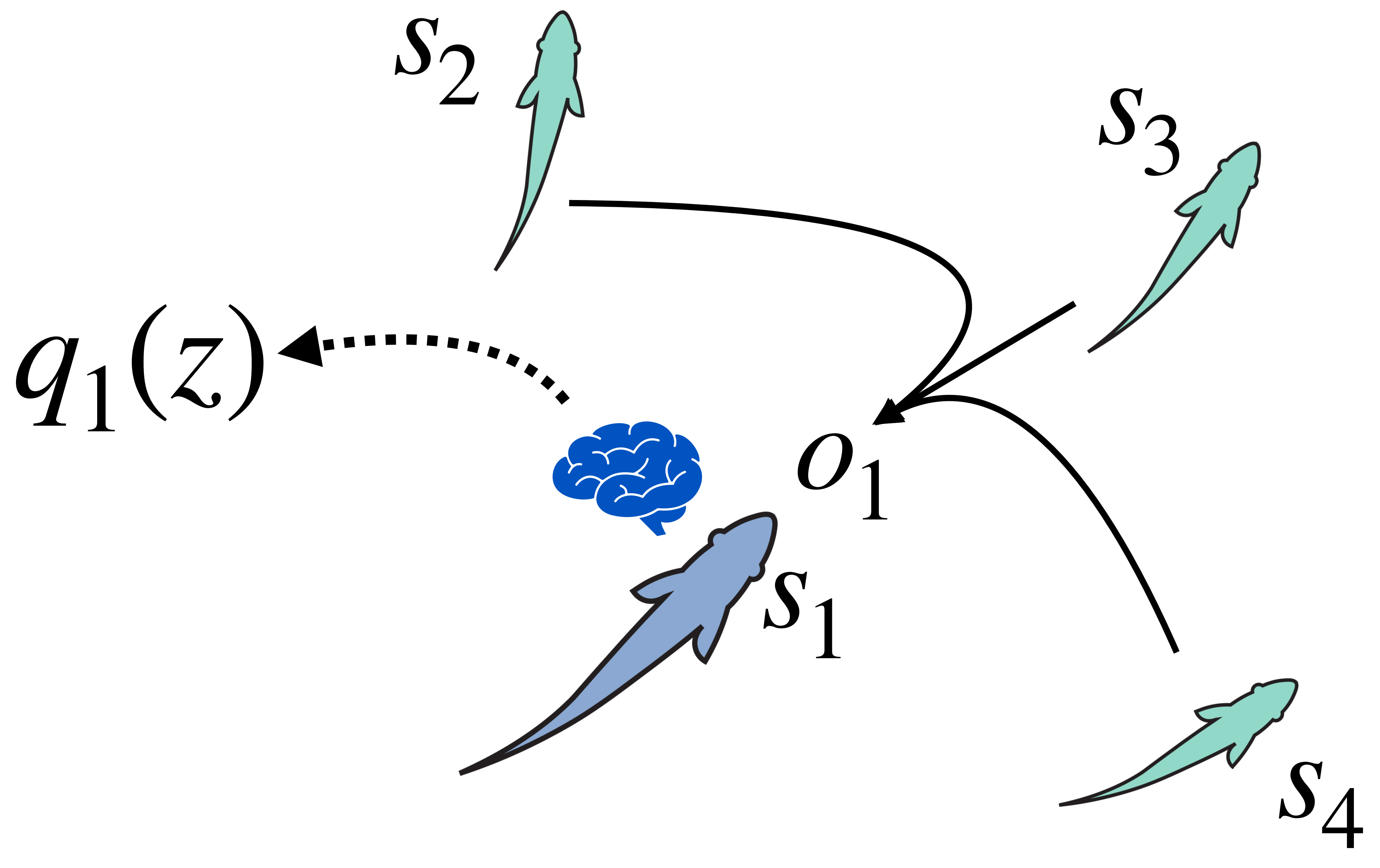
Potential function vs. free energy function(al)

Collective Bayesian (active inference) systems

- **Local functional** of probabilistic beliefs about one's environment
- Individual dynamics driven by dual gradient flows (action and perception) on this moving functional

$$F_1(o_1, q_1)$$

Free energy functional of probabilistic beliefs



$$\dot{q}_1 \propto -\nabla_{q_1} F_1$$

Perception

$$\dot{a}_1 \propto -\nabla_{a_1} F_1$$

Action

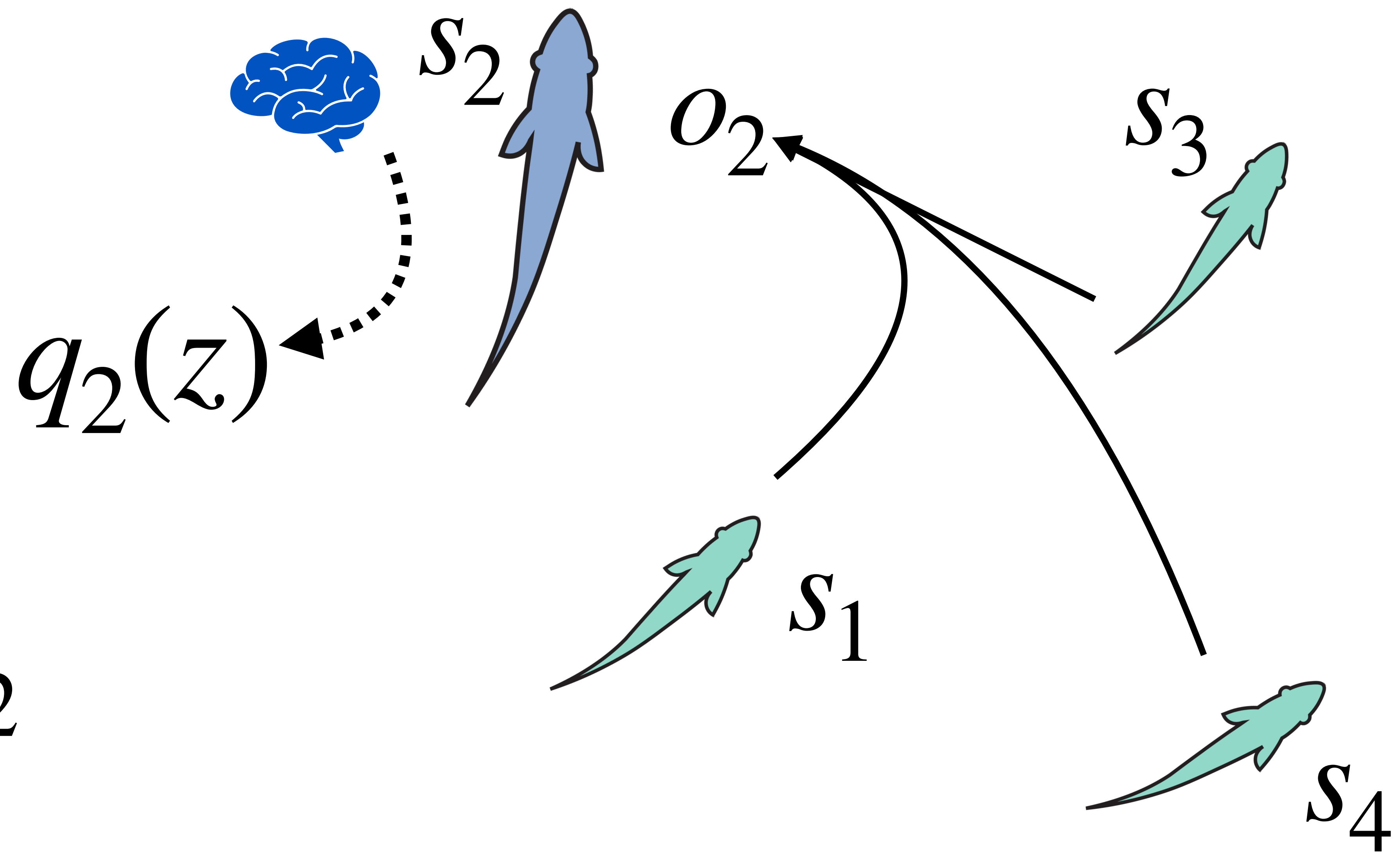
Potential function vs. free energy function(al)

Collective Bayesian (active inference) systems

- **Local functional** of probabilistic beliefs about one's environment
- Individual dynamics driven by dual gradient flows (action and perception) on this moving functional

$$F_2(o_2, q_2)$$

Free energy functional of probabilistic beliefs



$$\dot{q}_2 \propto -\nabla_{q_2} F_2$$

Perception

$$\dot{a}_2 \propto -\nabla_{a_2} F_2$$

Action

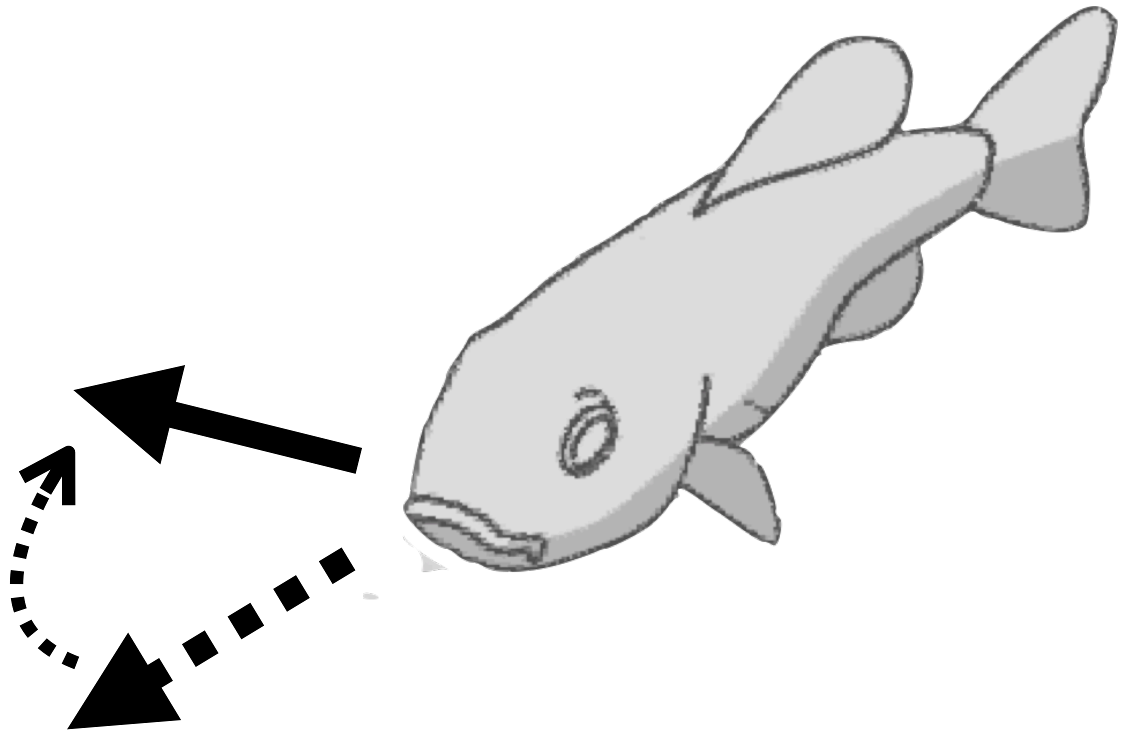
How do properties of individual models determine collective outcomes?

Modify the generative model of single agents and measure consequences

$$p(y, x) = p(y \mid x)p(x)$$

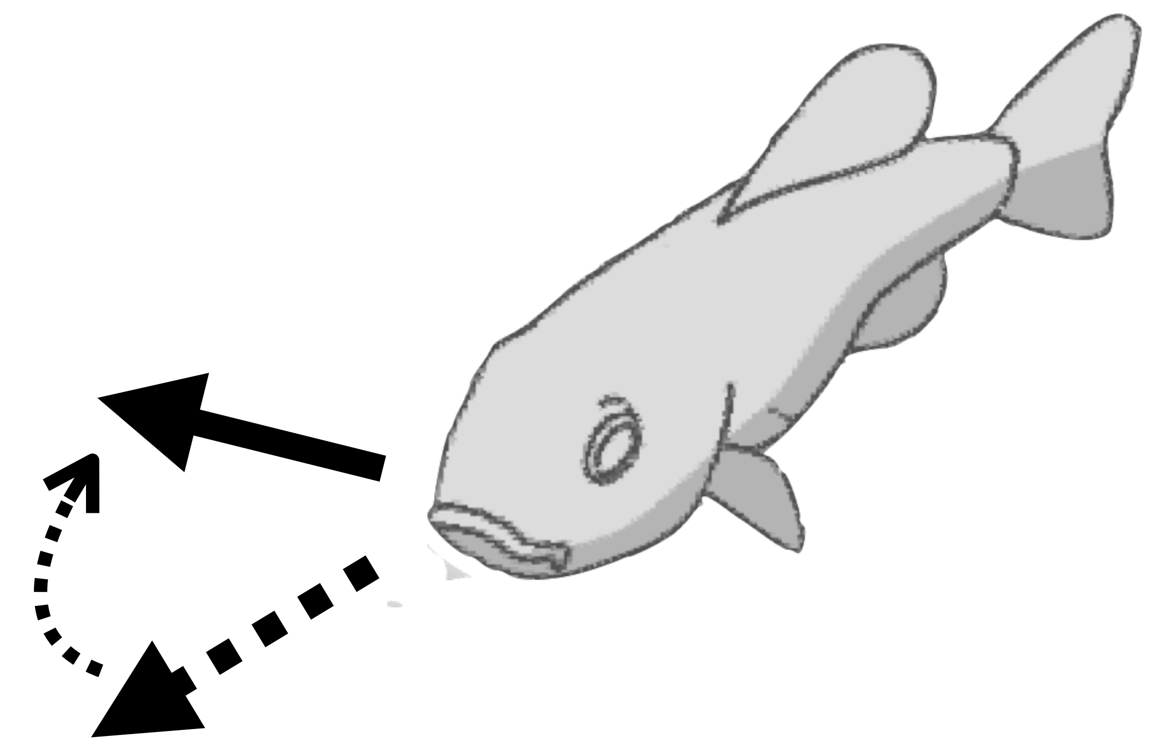
Weights vs beliefs

$$\text{Behavior} \quad = f_{social}(x) + f_{env}(z) + \epsilon$$



Weights vs beliefs

$$\text{Behavior} \quad = \omega_1 f_{social}(x) + \omega_2 f_{env}(z) + \epsilon$$

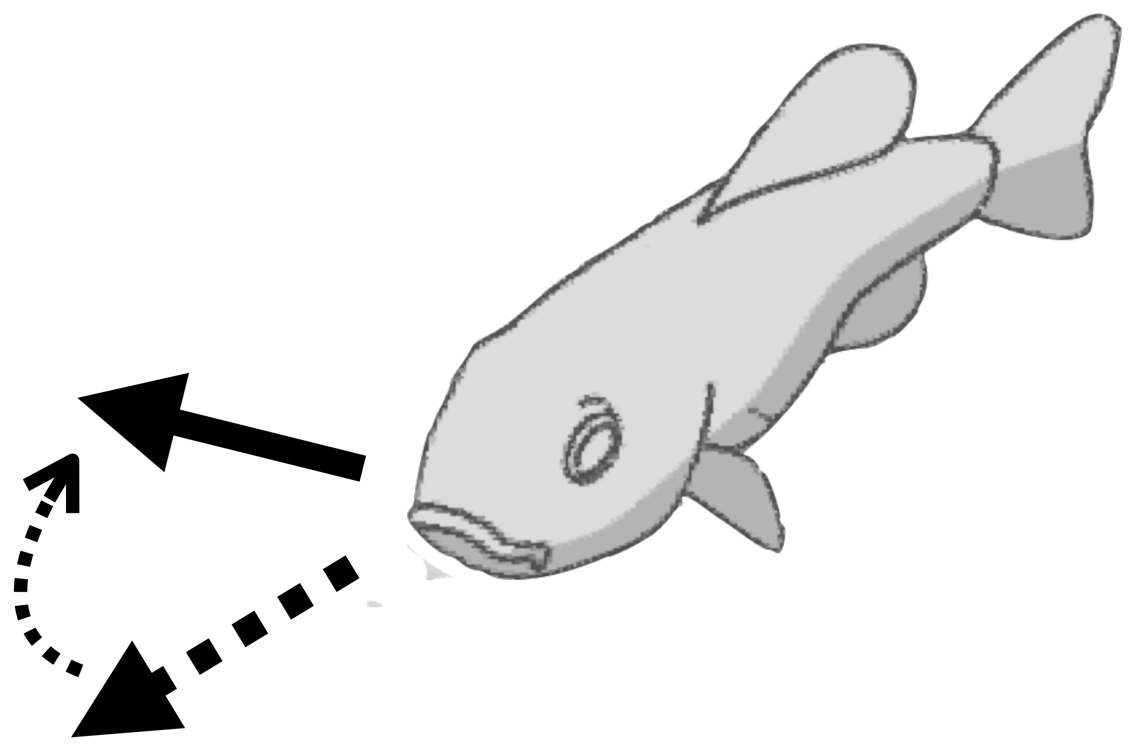


$$\omega_1, \omega_2 \quad ?$$

Weights vs beliefs

Sensory info 1 | Sensory info 2 | Prior info

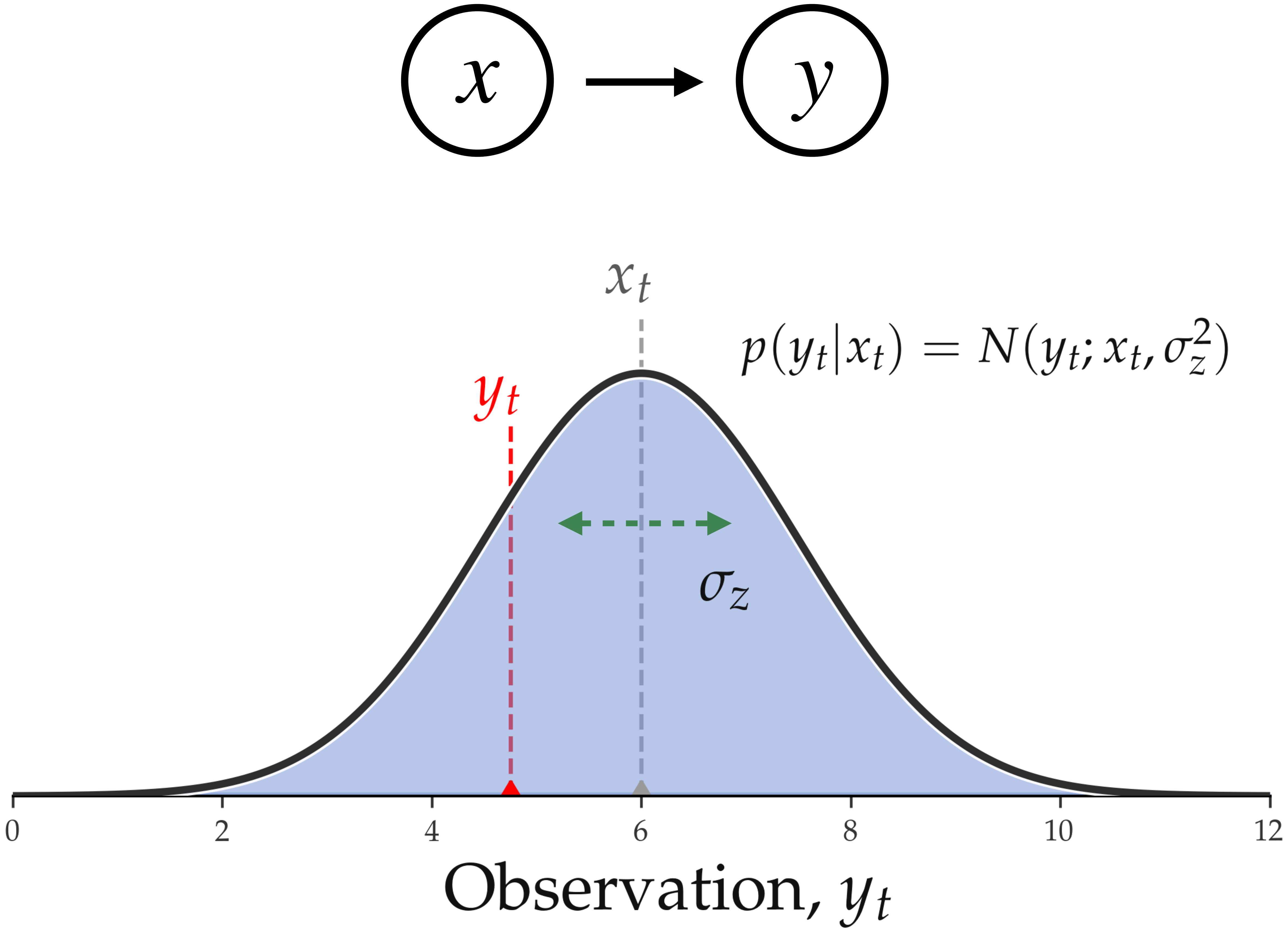
Behavior = E + E + P

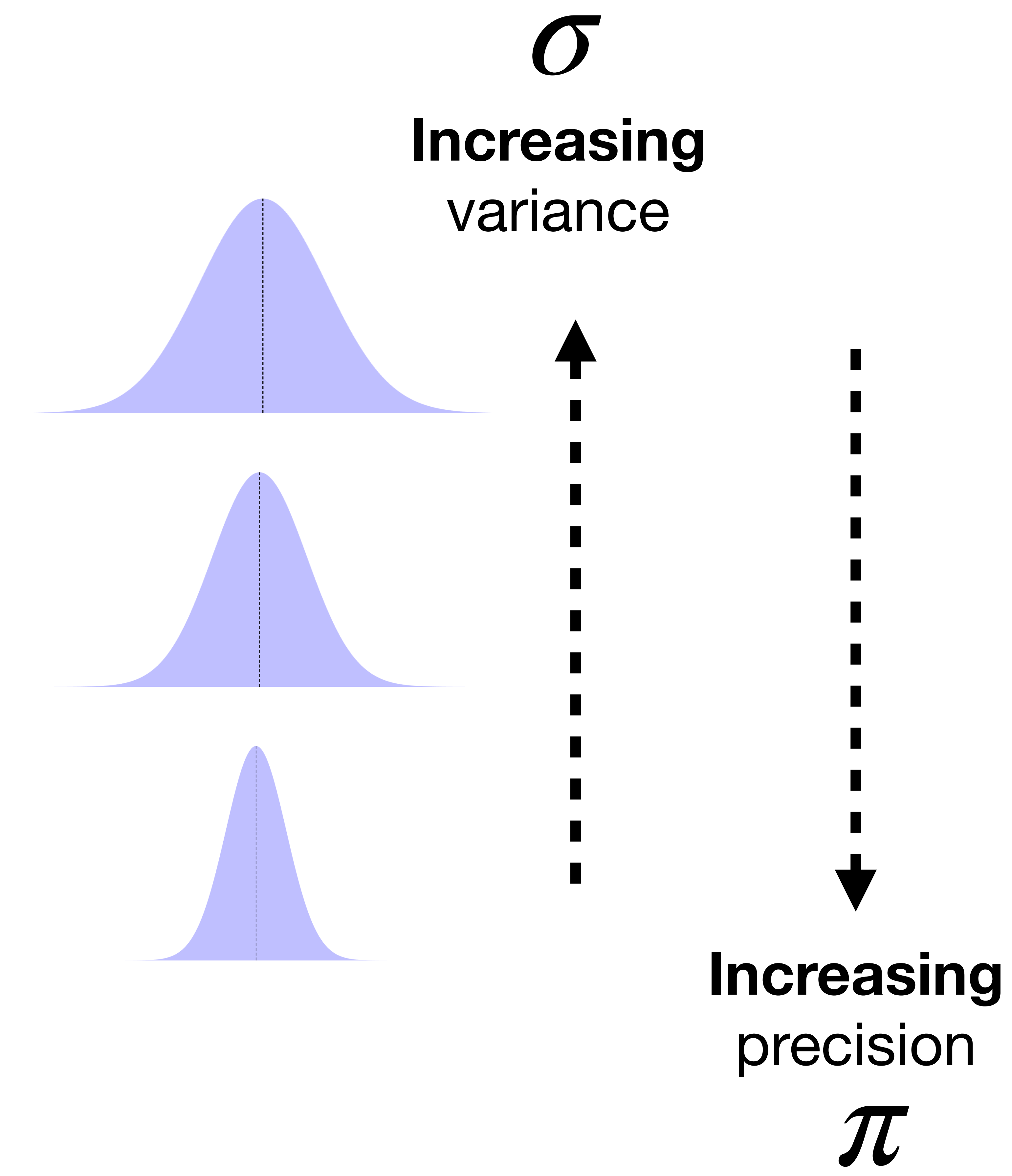


Beliefs about reliability of different types of sensory and prior information

Re-interpreting force-weights as beliefs about information reliability

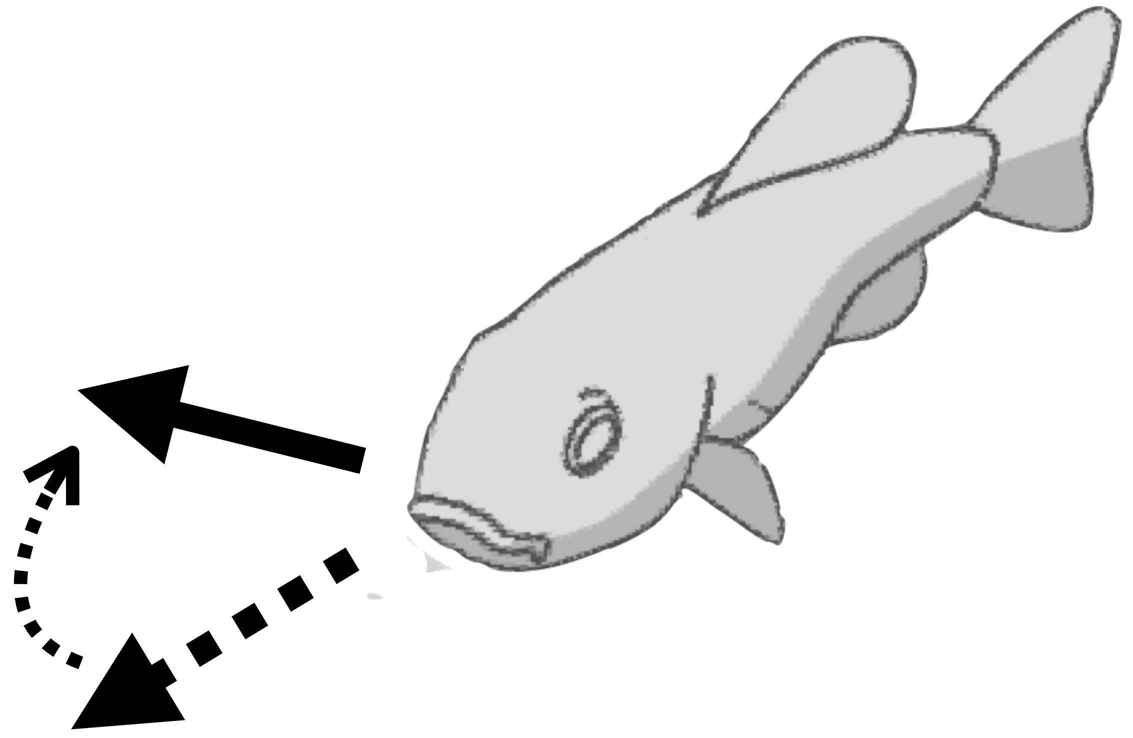
Information from different sources can be more or less "trustworthy"





Re-interpreting force-weights as beliefs about sensory reliability

$$\text{Movement} \;= \omega_1 f_{social}(x) + \omega_2 f_{env}(z) + \epsilon$$

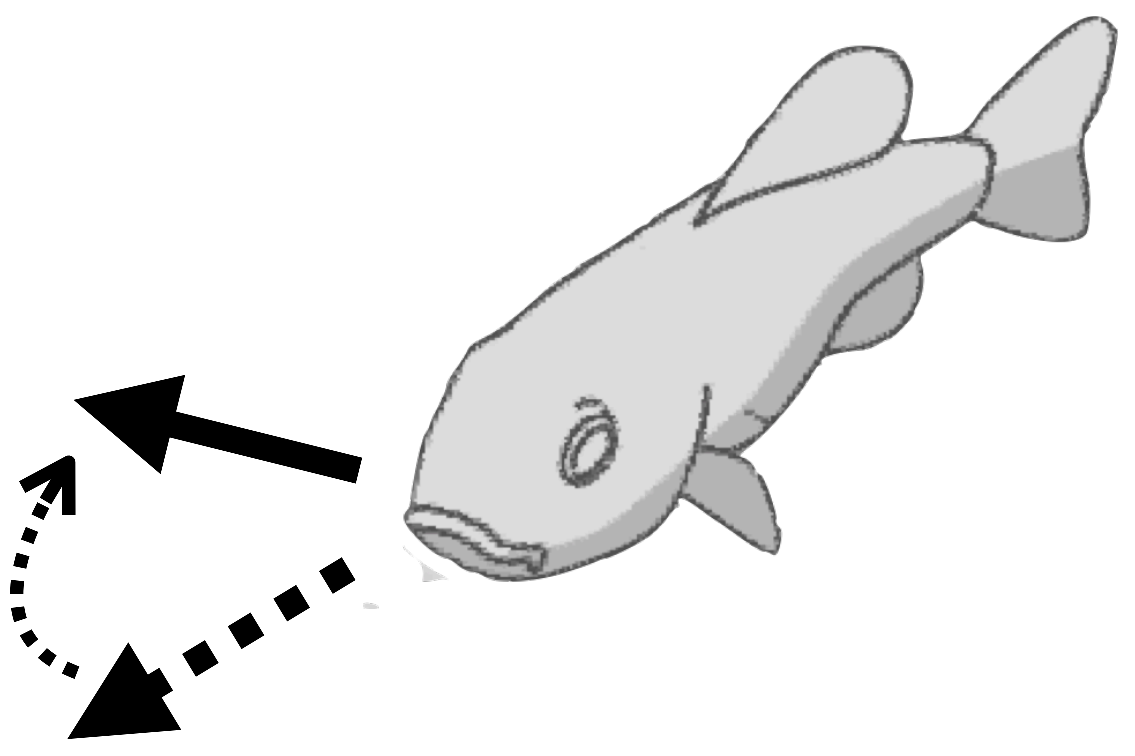


ω_1, ω_2 ?

Re-interpreting force-weights as beliefs about sensory reliability

Prediction errors

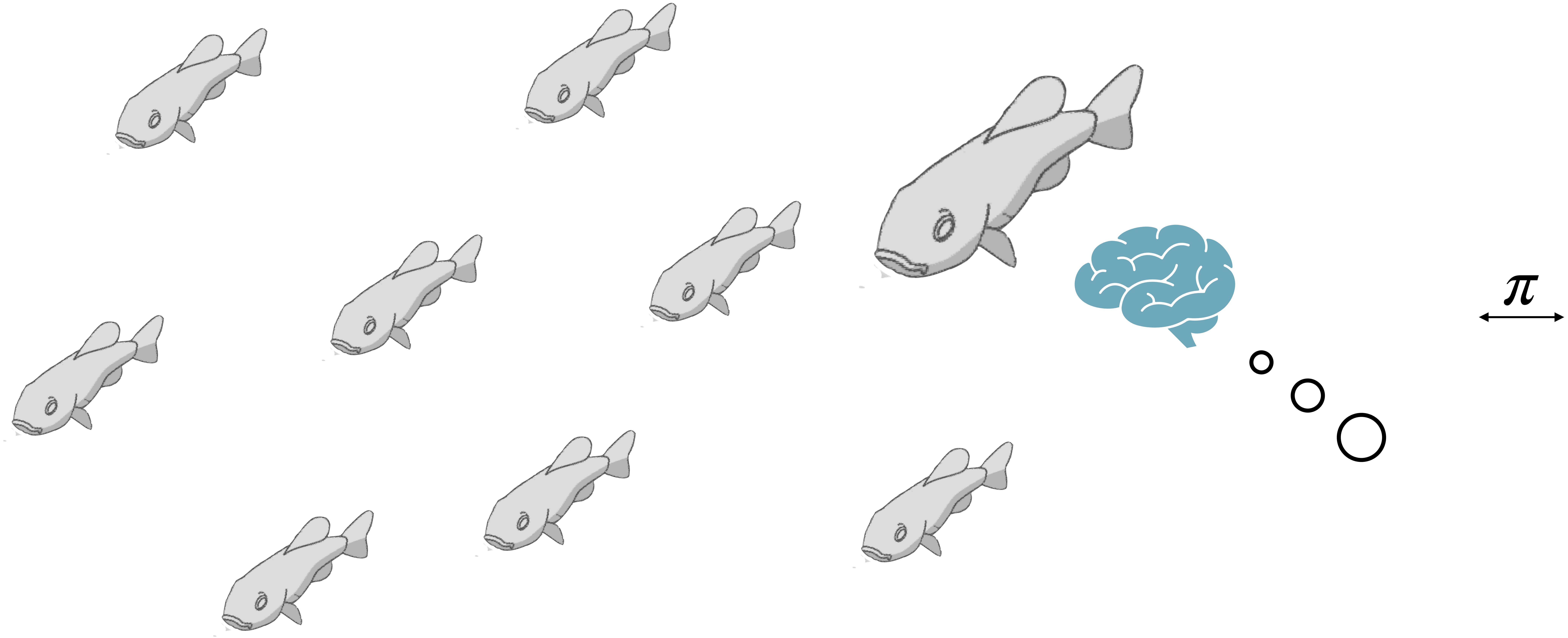
$$\text{Movement} = \pi_1 \boxed{\epsilon_1} + \pi_2 \boxed{\epsilon_2}$$



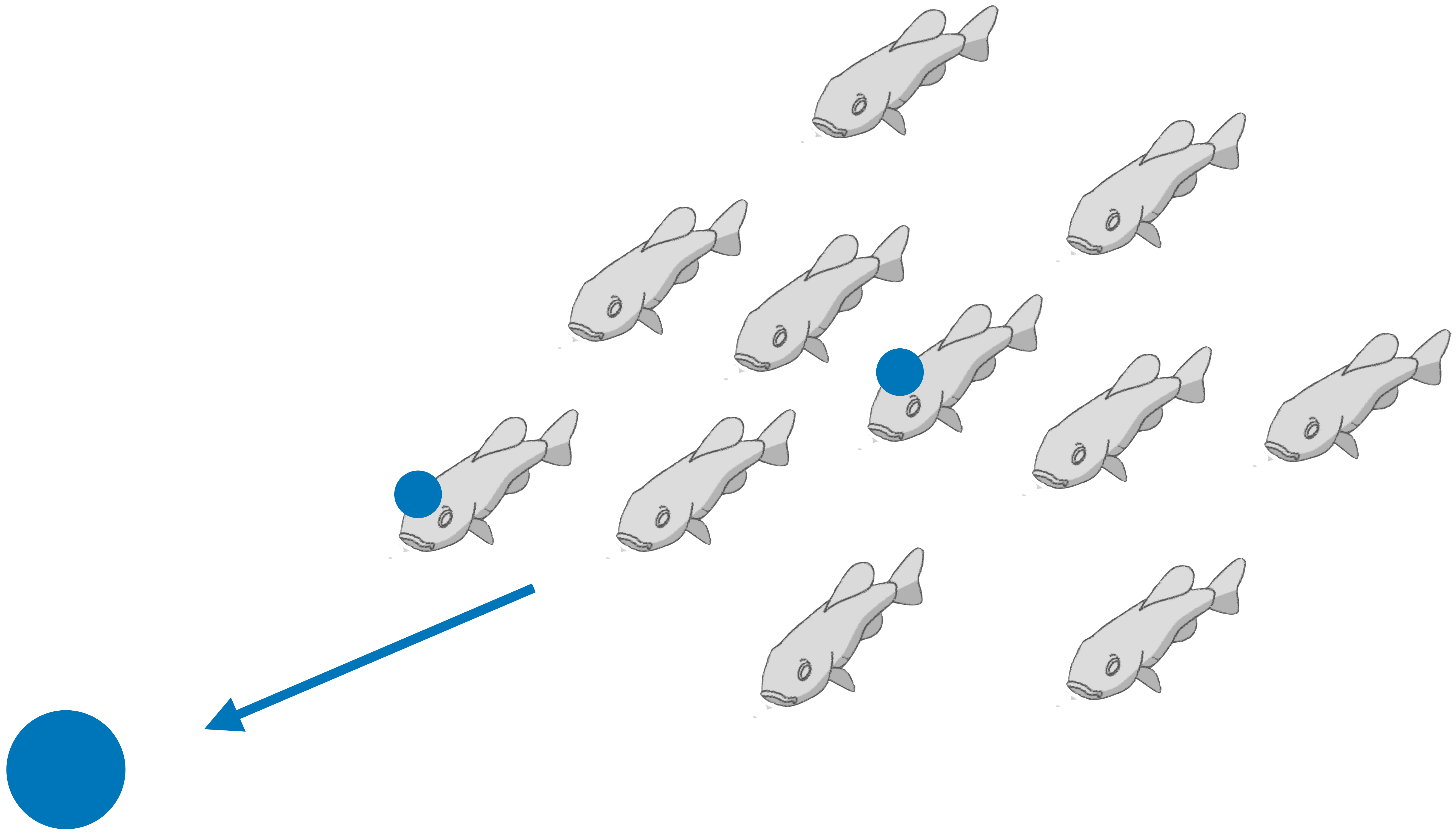
π_1 Beliefs about reliability of signal 1

π_2 Beliefs about reliability of signal 2

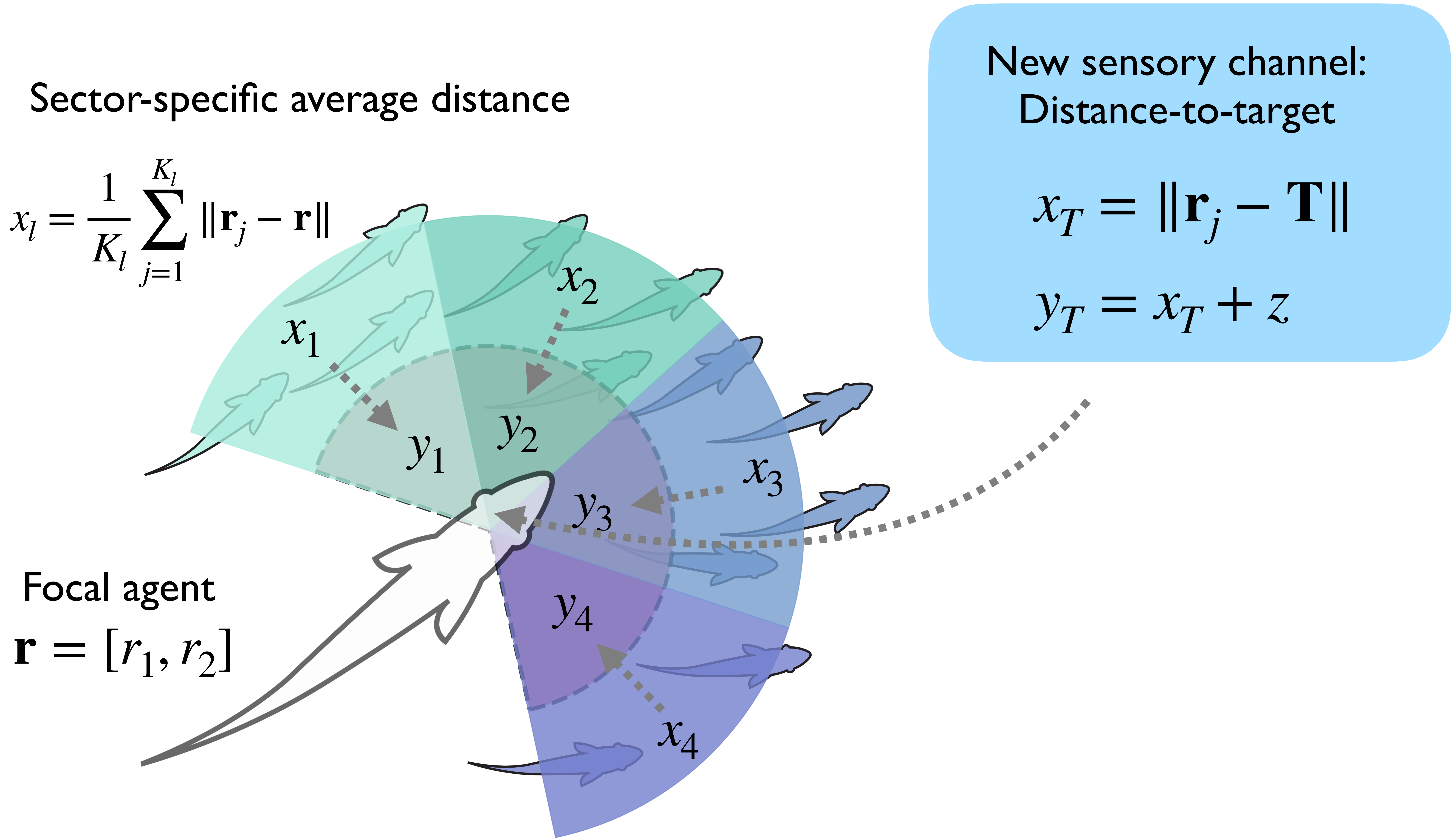
How do individual beliefs determine collective information processing?



Collective information transfer

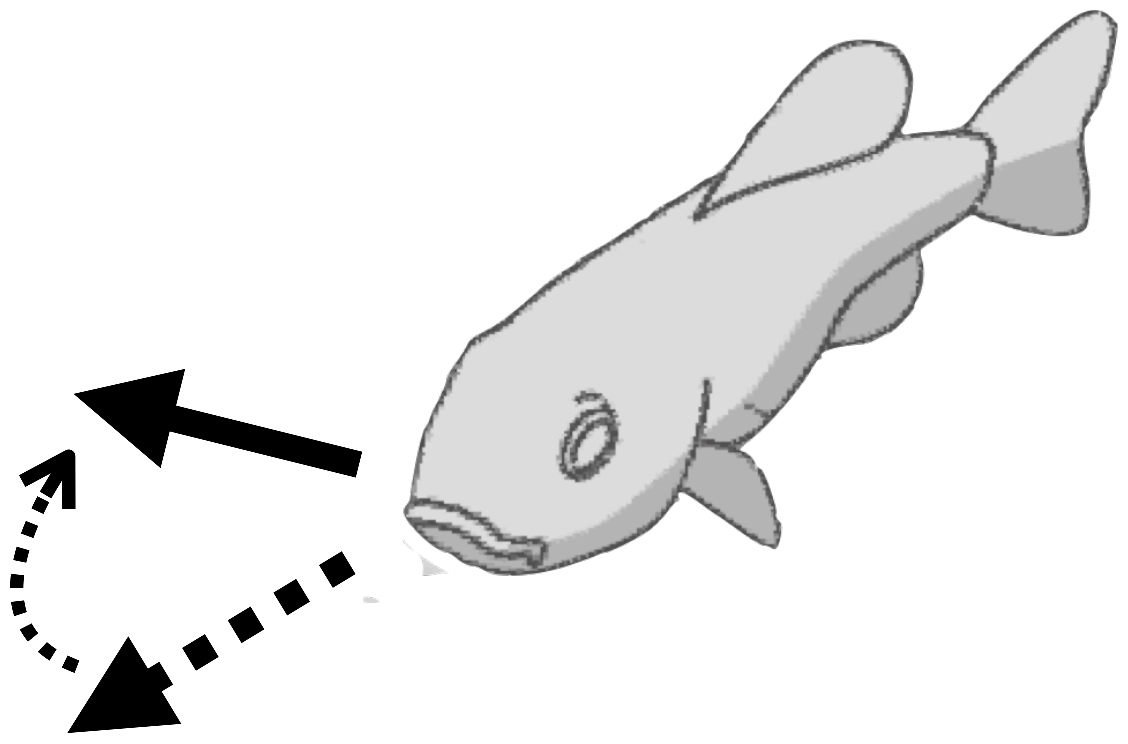


Some agents have an extra source of sensory information



Action becomes a precision-weighted sum of vectors

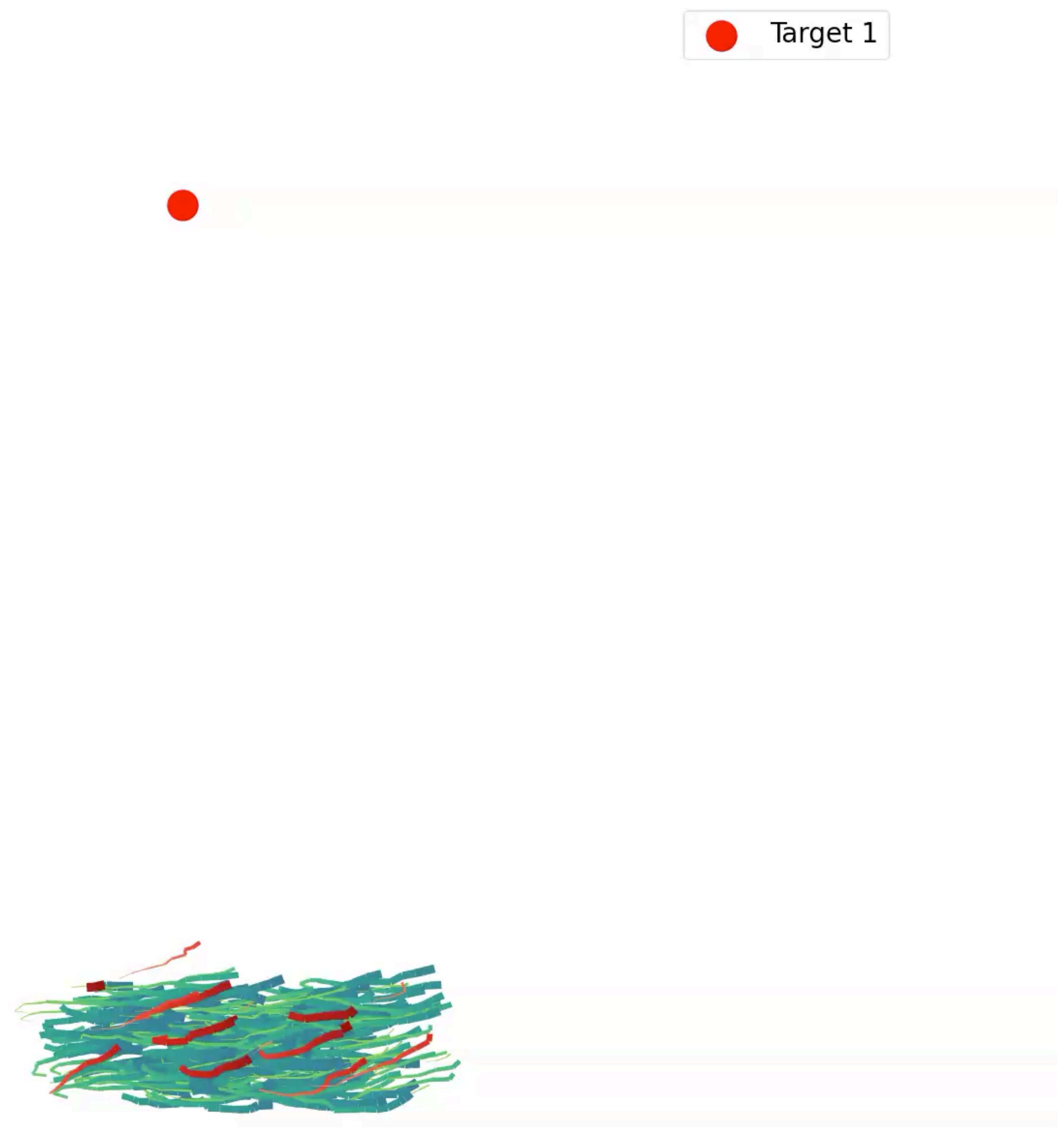
$$\text{Heading direction} = \pi_{Soc} \times \uparrow + \pi_{Target} \times \nearrow$$



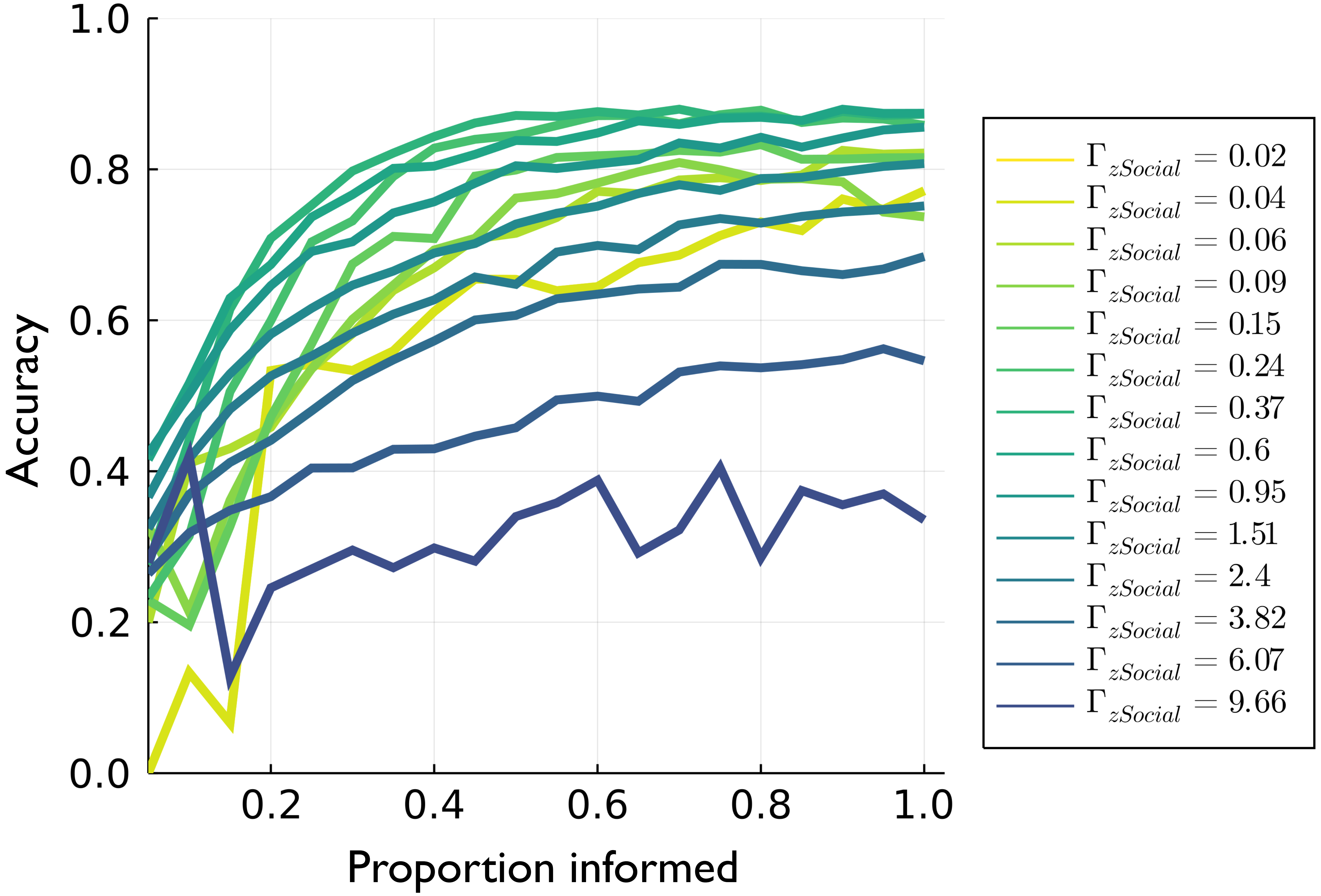
π_{Soc} Beliefs about reliability of social info

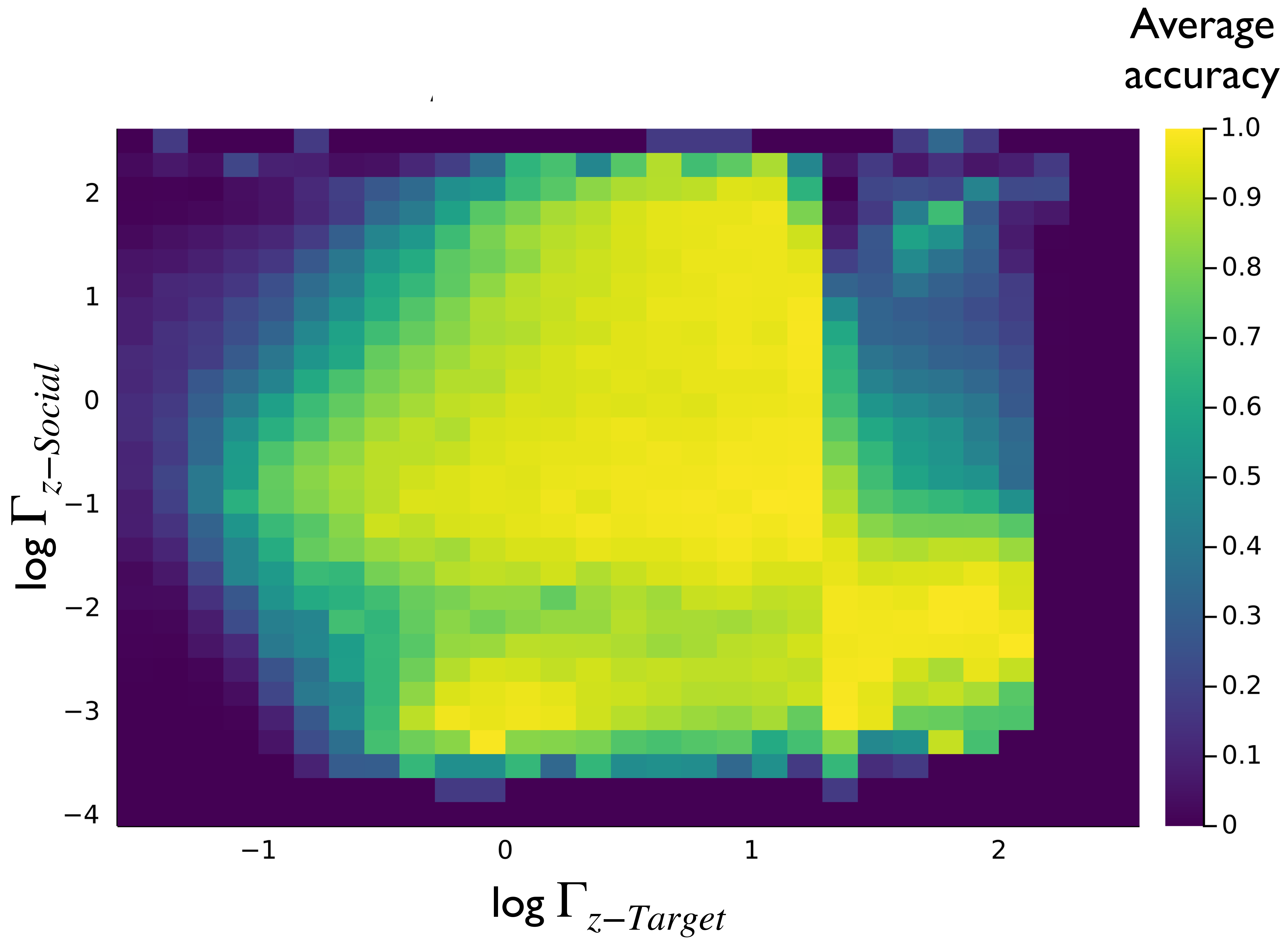
π_{Target} Beliefs about reliability of target info

Collective navigates to target despite uninformed individuals



High accuracy within a large regime of social and target precisions

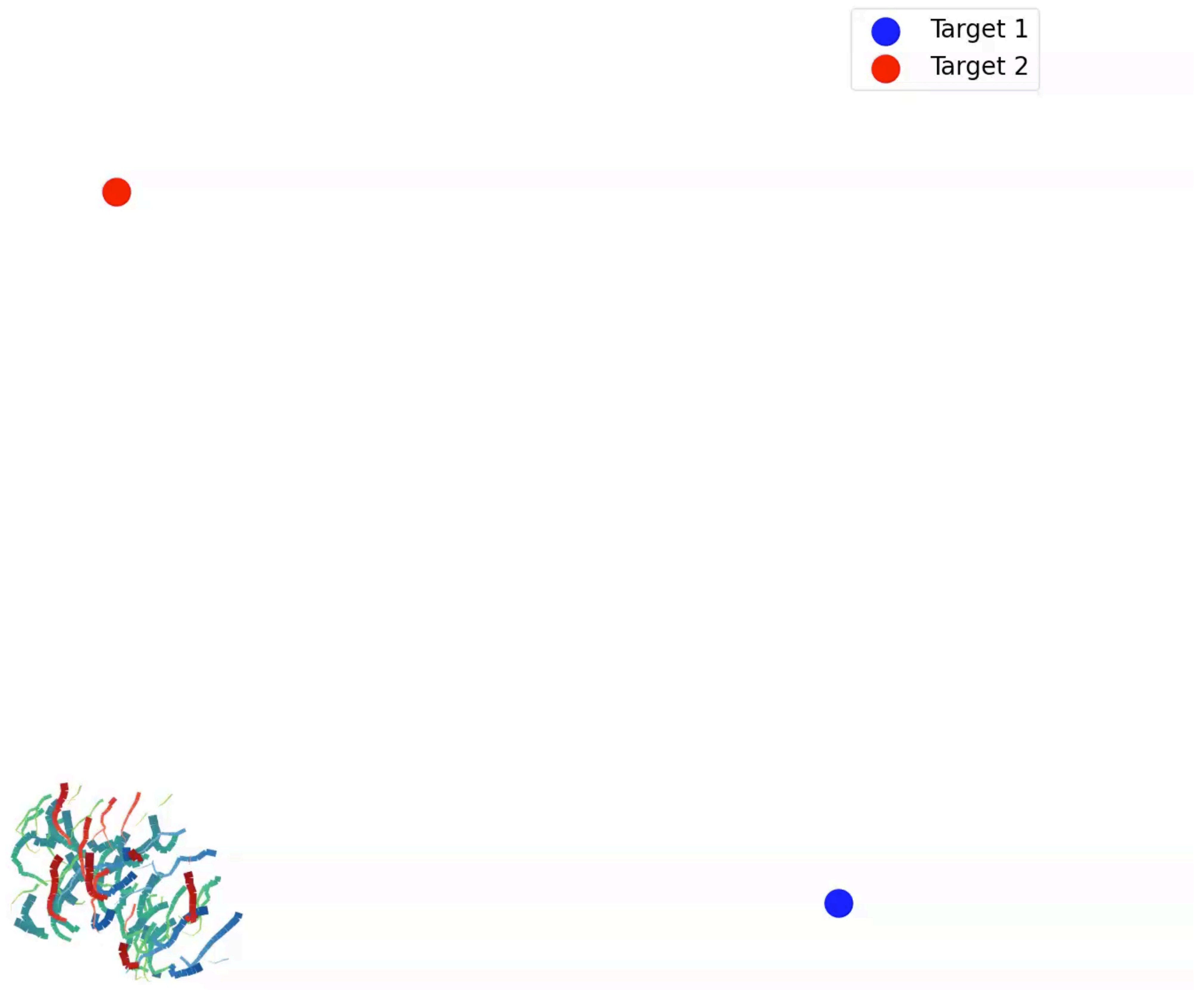




$$\Gamma_{z-Social} = \pi_{Social} \qquad \Gamma_{z-Target} = \pi_{Target}$$

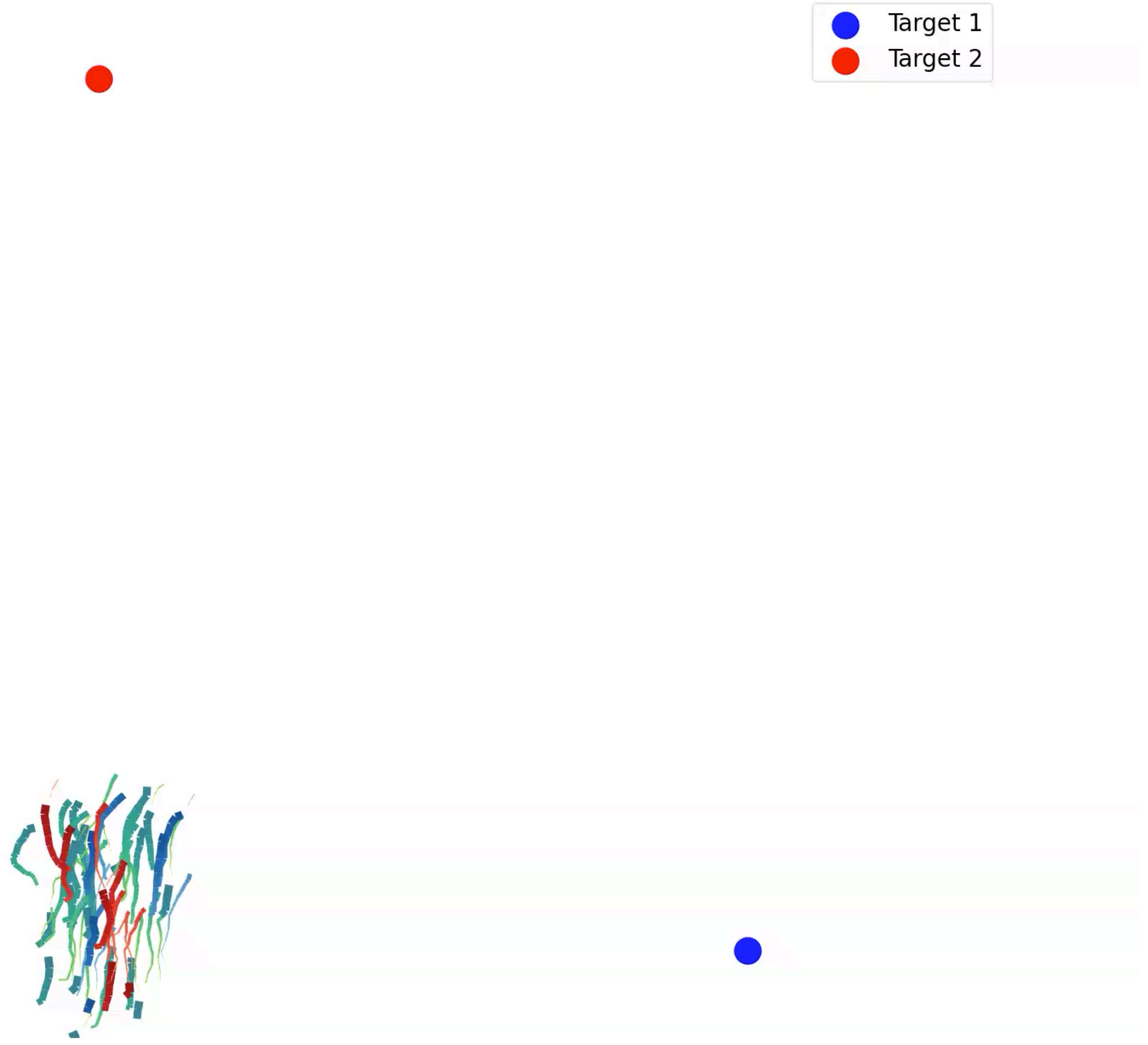
Group fission in presence of multiple targets

$$\Gamma_{z-Social} = \pi_{Social} = 1$$



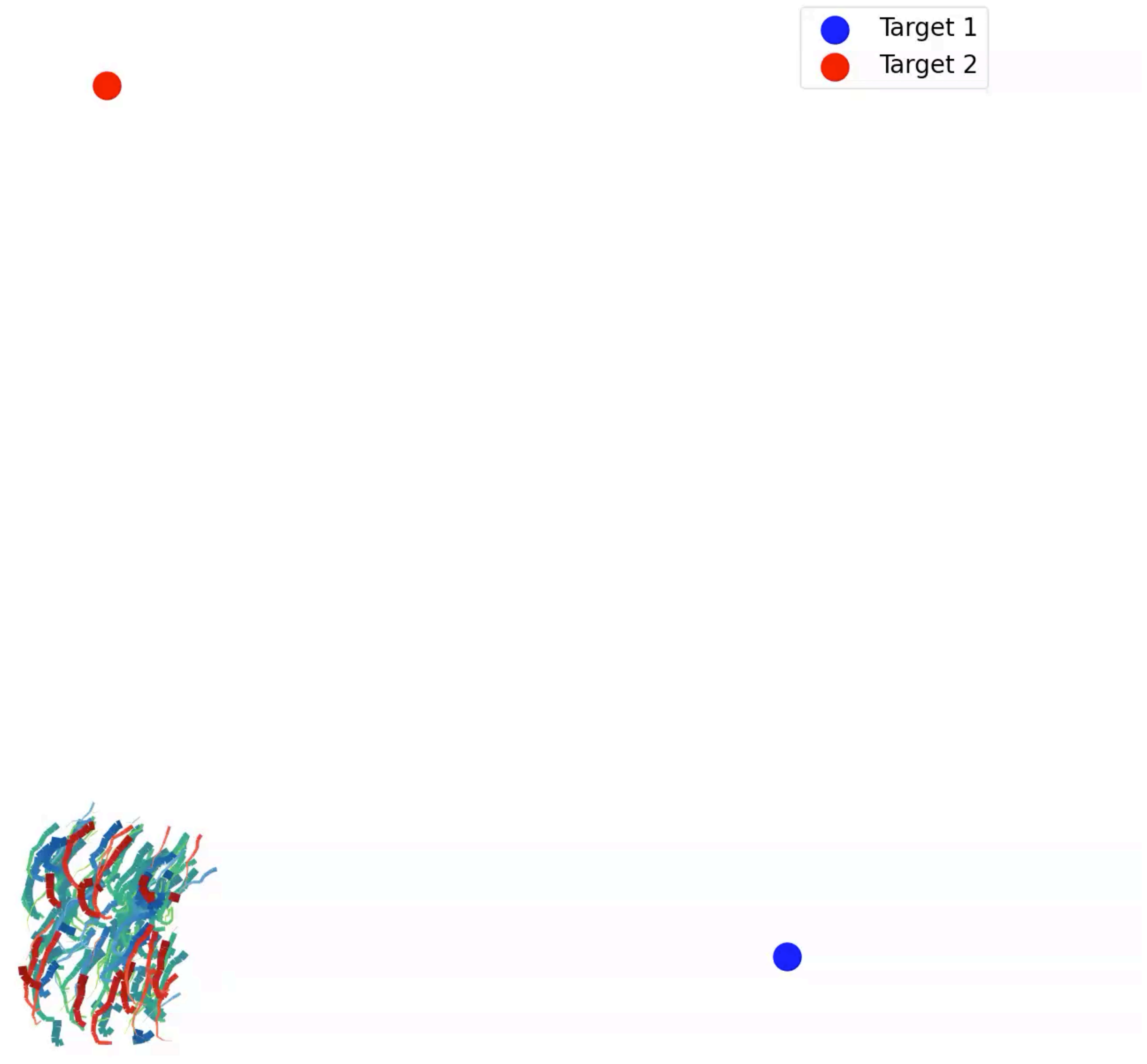
Decreasing social precision leads to consensus

$$\Gamma_{z-Social} = \pi_{Social} = \frac{1}{4}$$



Oscillations between the targets

$$\Gamma_{z-Social} = \pi_{Social} = \frac{1}{4}$$



Adapting the generative model

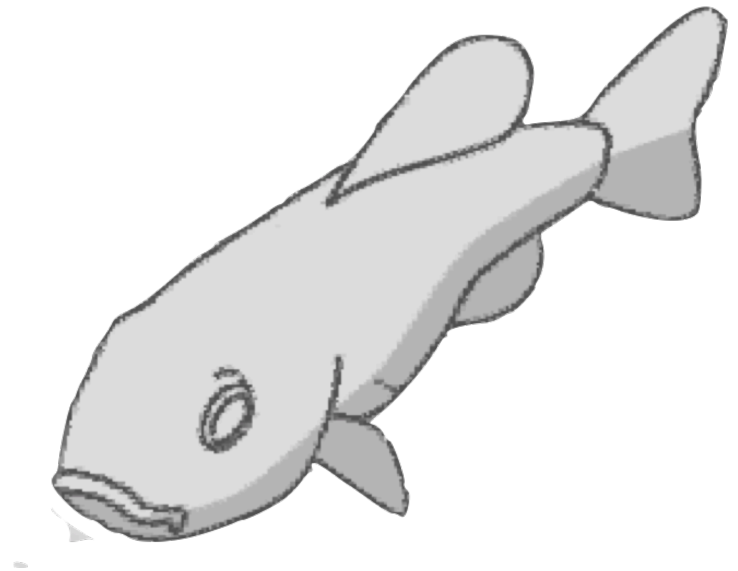
$F =$ Surprise $=$ Prediction error

$$\frac{d\mu}{dt} = -\nabla_{\mu} F(\mu, y)$$

Perception

$$\frac{d\mathbf{v}}{dt} = -\nabla_{\mathbf{v}} F(\mu, y(\mathbf{v}))$$

Action

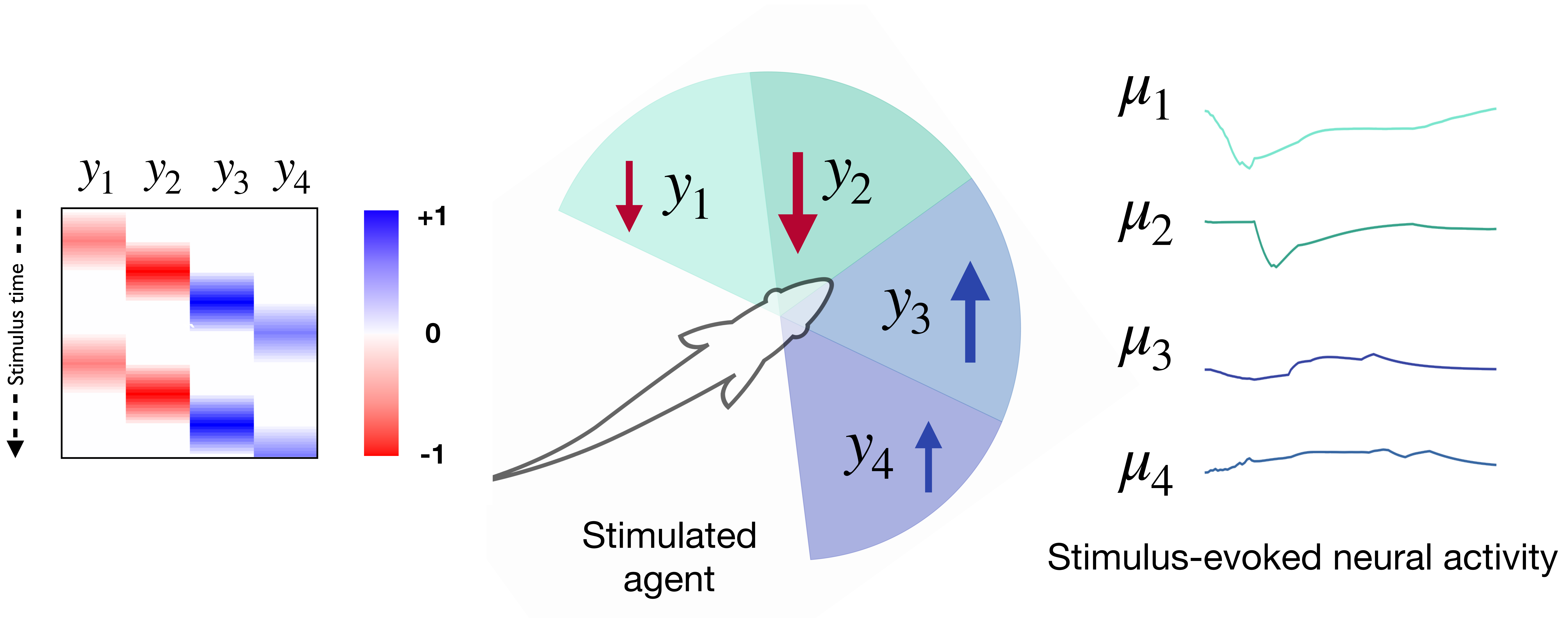


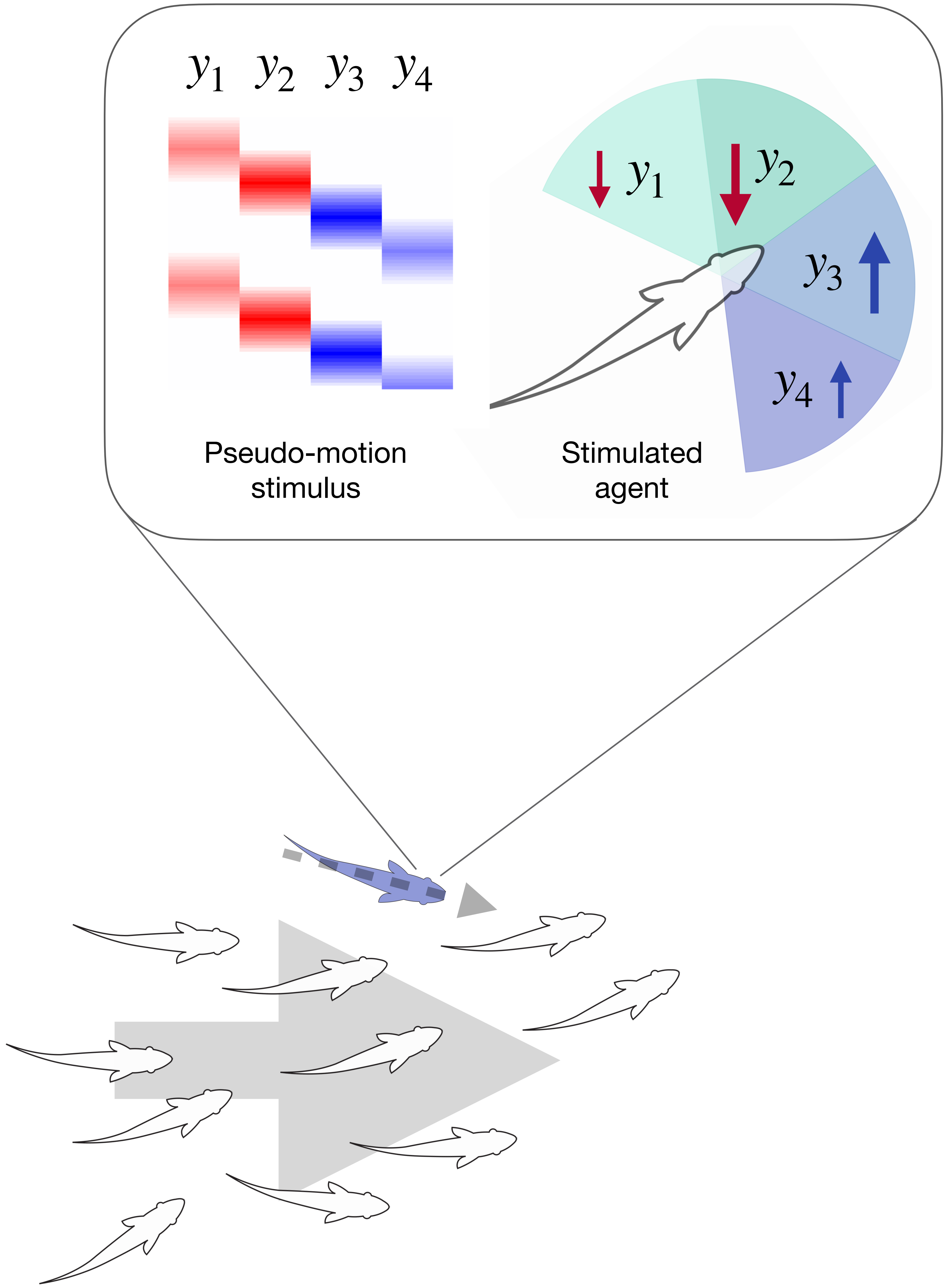
π_z

$$\frac{d\theta}{dt} = -\nabla_{\theta} F(\mu, y, \theta)$$

Plasticity

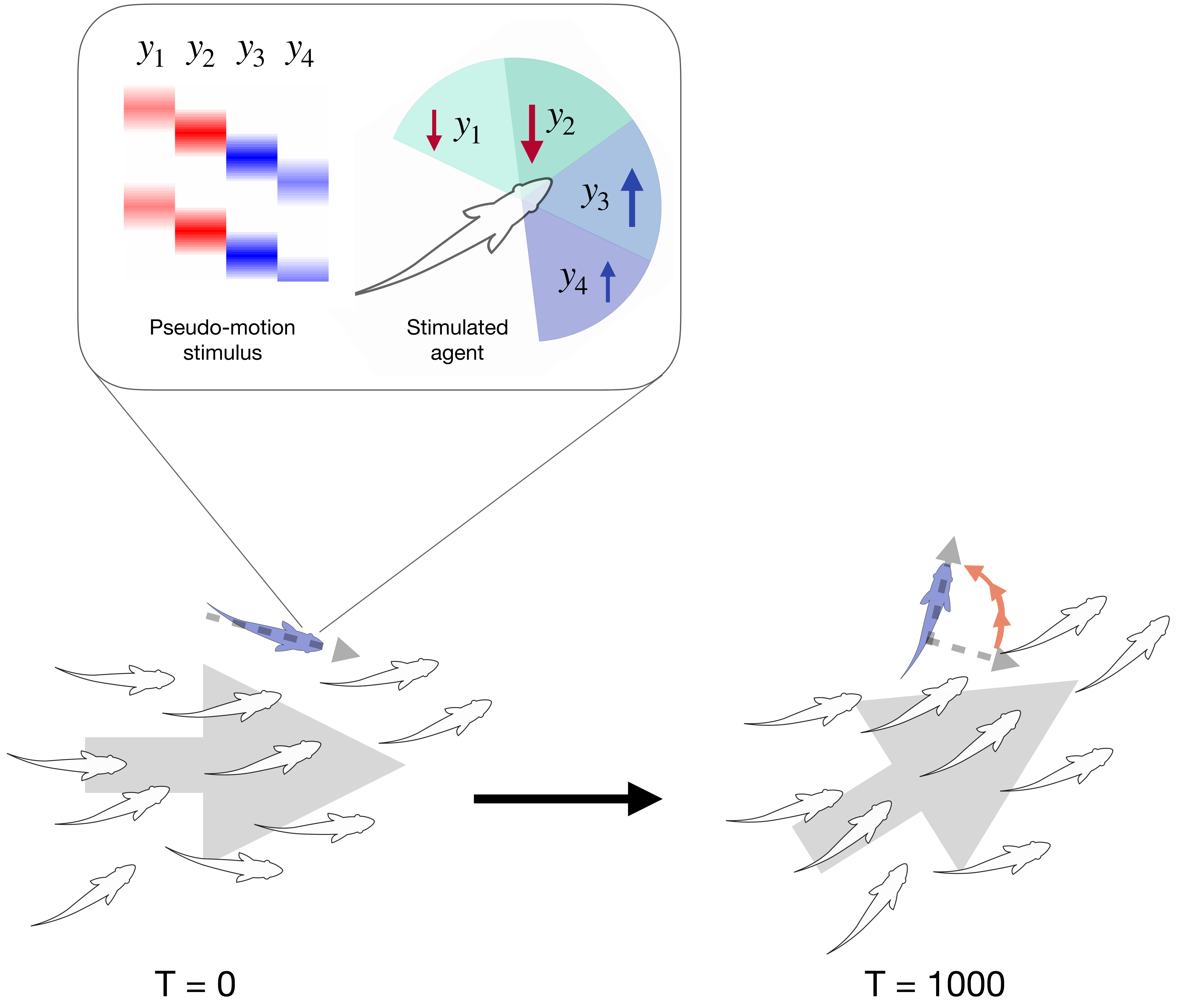
Phantom prediction errors in single agents



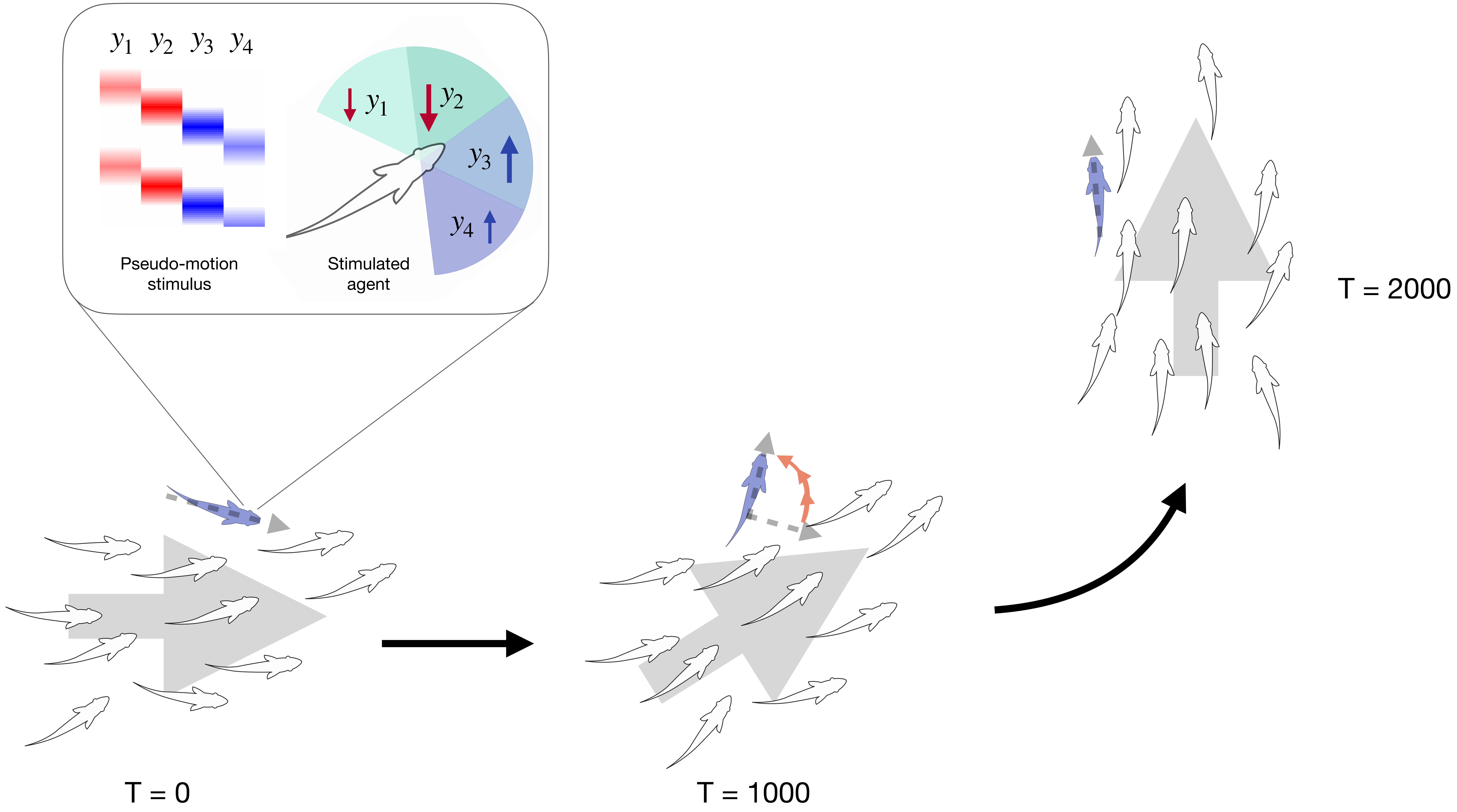


y_1 y_2 y_3 y_4
Pseudo-motion stimulus
y_1
y_2
y_3
y_4
Stimulated agent

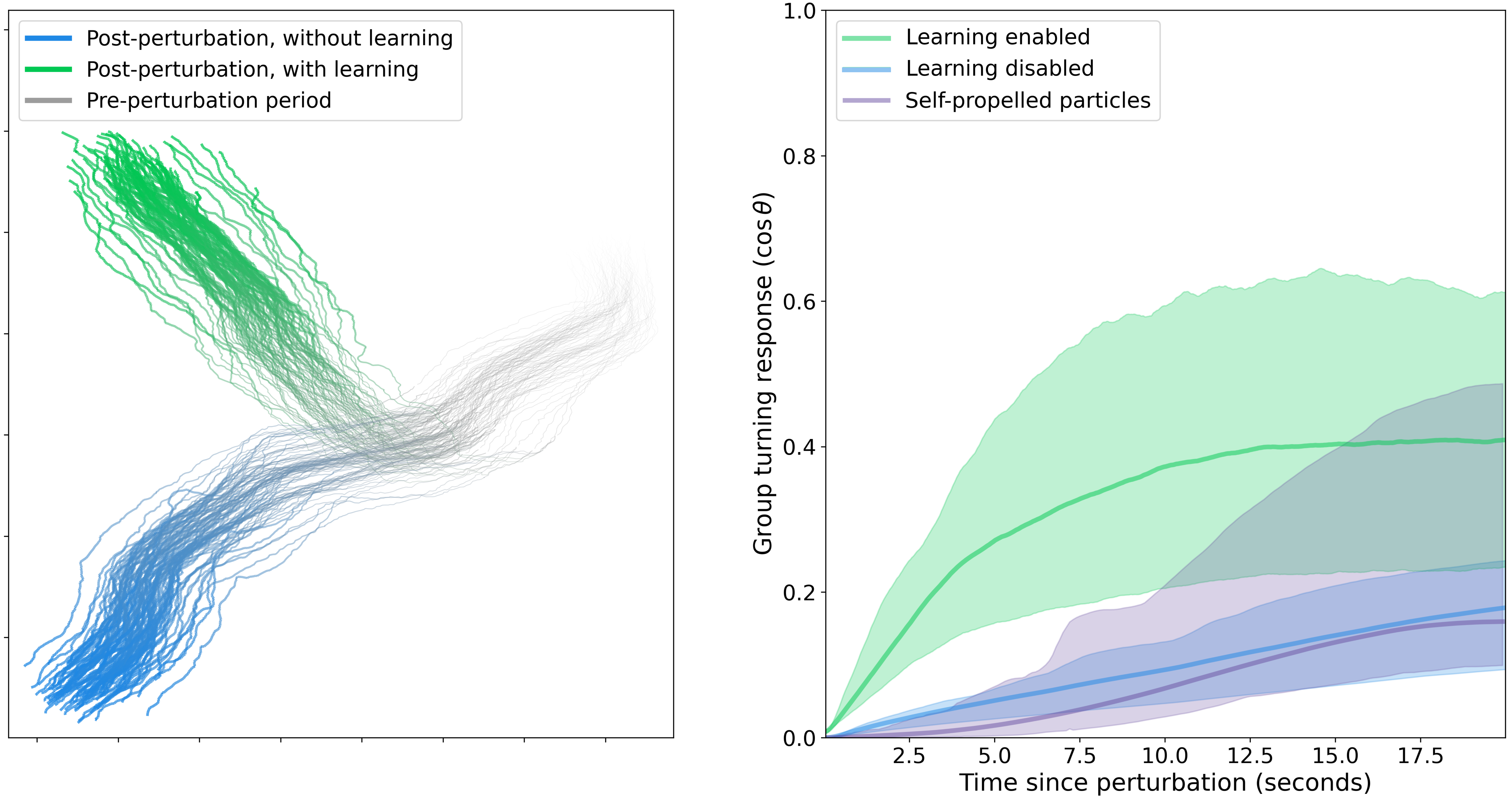
T = 0

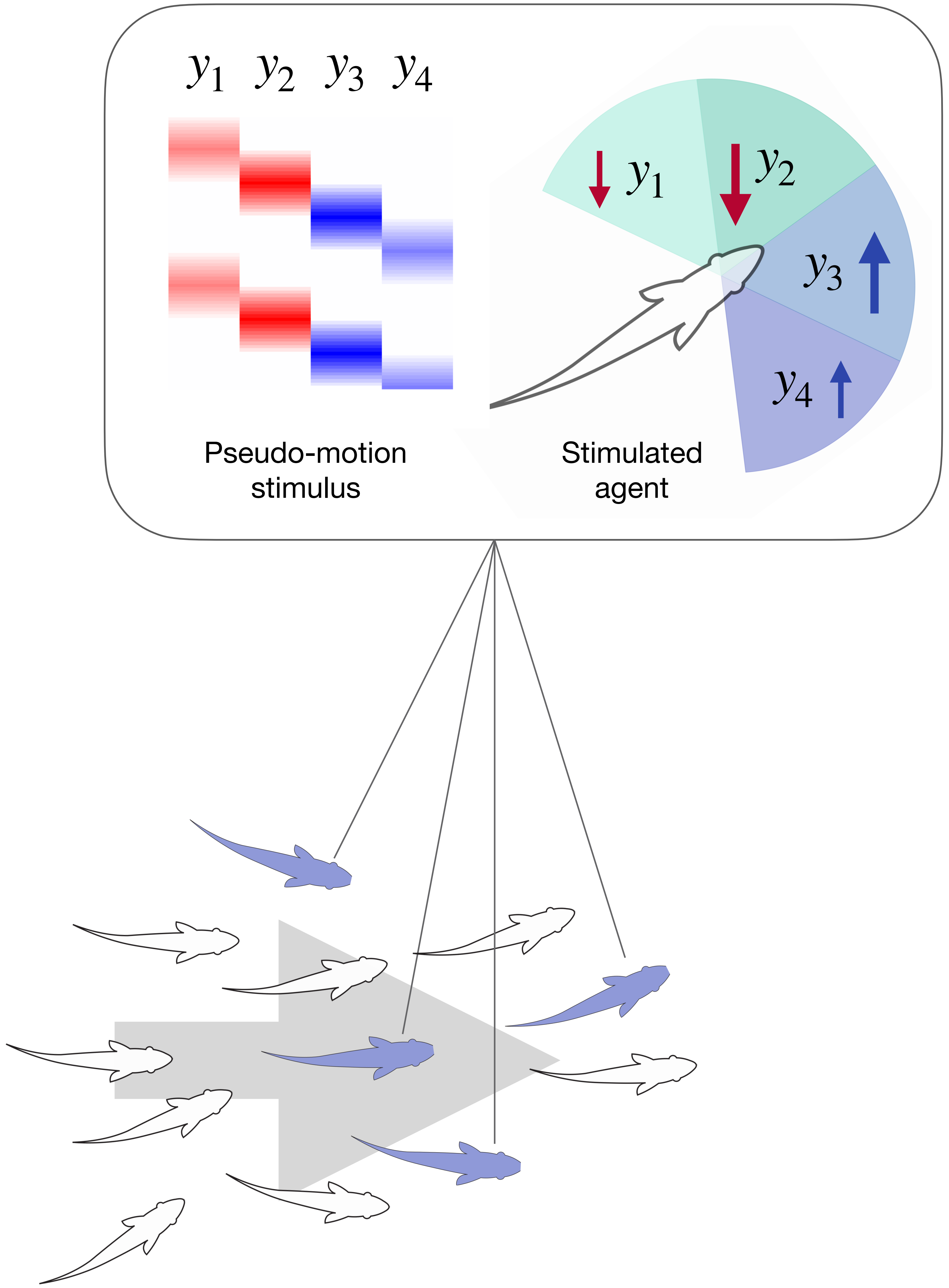


y_1 y_2 y_3 y_4
y_1
y_2
y_3
y_4
Pseudo-motion stimulus
Stimulated agent
T = 0
T = 1000

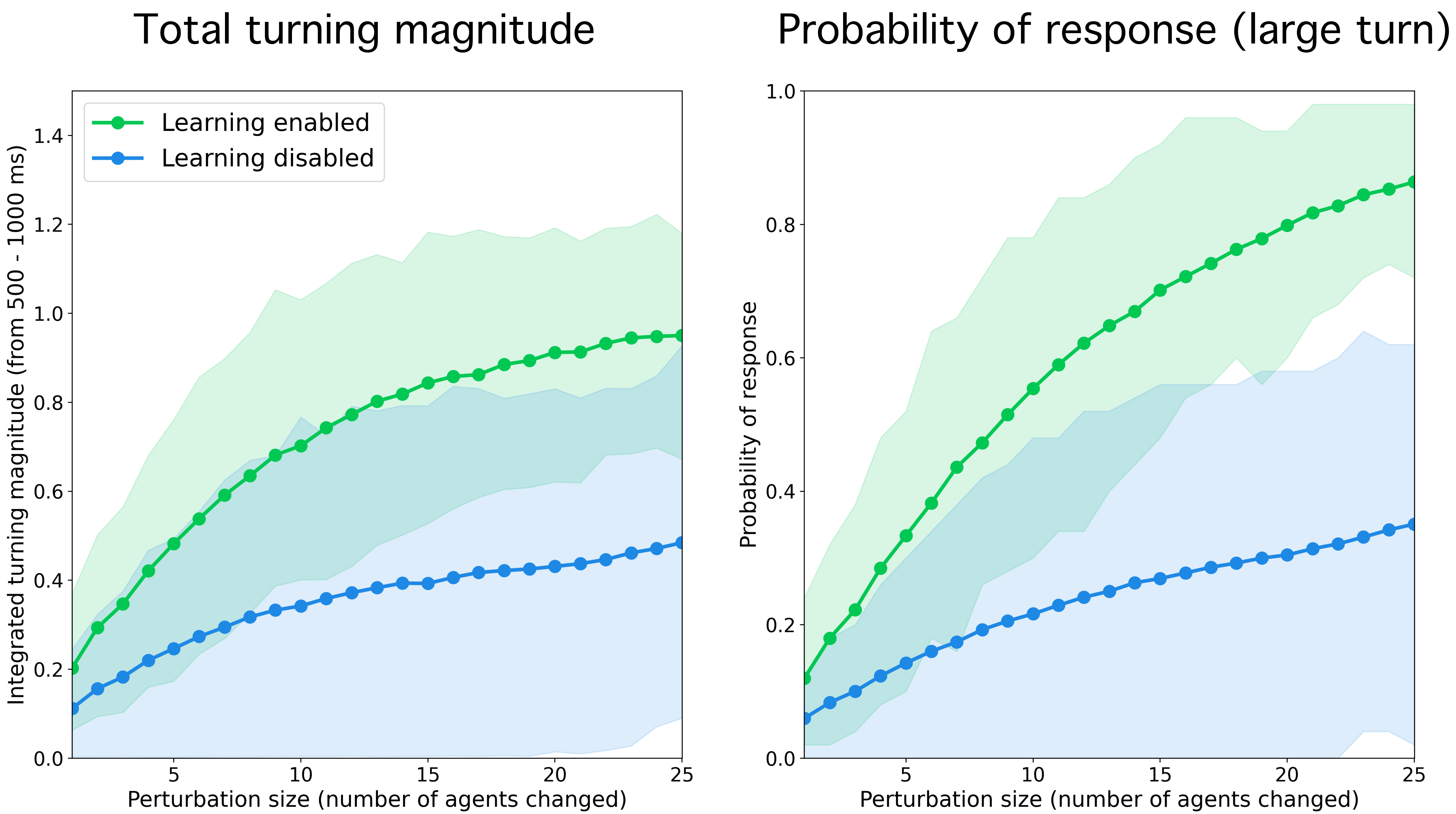


y_1 y_2 y_3 y_4
y_1
y_2
y_3
y_4
Pseudo-motion stimulus
Stimulated agent
T = 0
T = 1000
T = 2000


Post-perturbation, without learning
Post-perturbation, with learning
Pre-perturbation period
Learning enabled
Learning disabled
Self-propelled particles
Group turning response ($\cos\theta$)
Time since perturbation (seconds)
0.0
0.2
0.4
0.6
0.8
1.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5



Vary perturbation size and measure response


Total turning magnitude
Probability of response (large turn)
Learning enabled
Learning disabled
Integrated turning magnitude (from 500 - 1000 ms)
Probability of response
Perturbation size (number of agents changed)
Perturbation size (number of agents changed)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
5
10
15
20
25

Conclusion

Collective behavior from surprise minimization

Conor Heins[a,b,c,e,1], Beren Millidge[d], Lancelot Da Costa[e,f,g], Richard P. Mann[h], Karl J. Friston[e,g], and Iain D. Couzin[a,b,c]

Thanks to the team behind this work

Iain Couzin

Karl Friston

Lance Da Costa

Beren Millidge

Richard Mann

Bayesian Machine Learning @ VERSES

- Scaling active inference and Bayesian neural networks to modern machine learning contexts
- Variational Bayes for Mixture Models
- Can we train deep nets faster, with less data, while also being Bayesian (i.e., quantifying uncertainty)?
- (Transformer) Attention as Inference

VERSES

Come chat to me if you want to learn more!

The goal of model-building

y	Data
θ	Parameters for model m
m	Model

At the end of the day (in my opinion),
science is about maximizing **model evidence**

$$p(y \mid m) = \sum_{\theta} p(y \mid \theta, m) p(\theta \mid m)$$

Also known as "marginal likelihood"

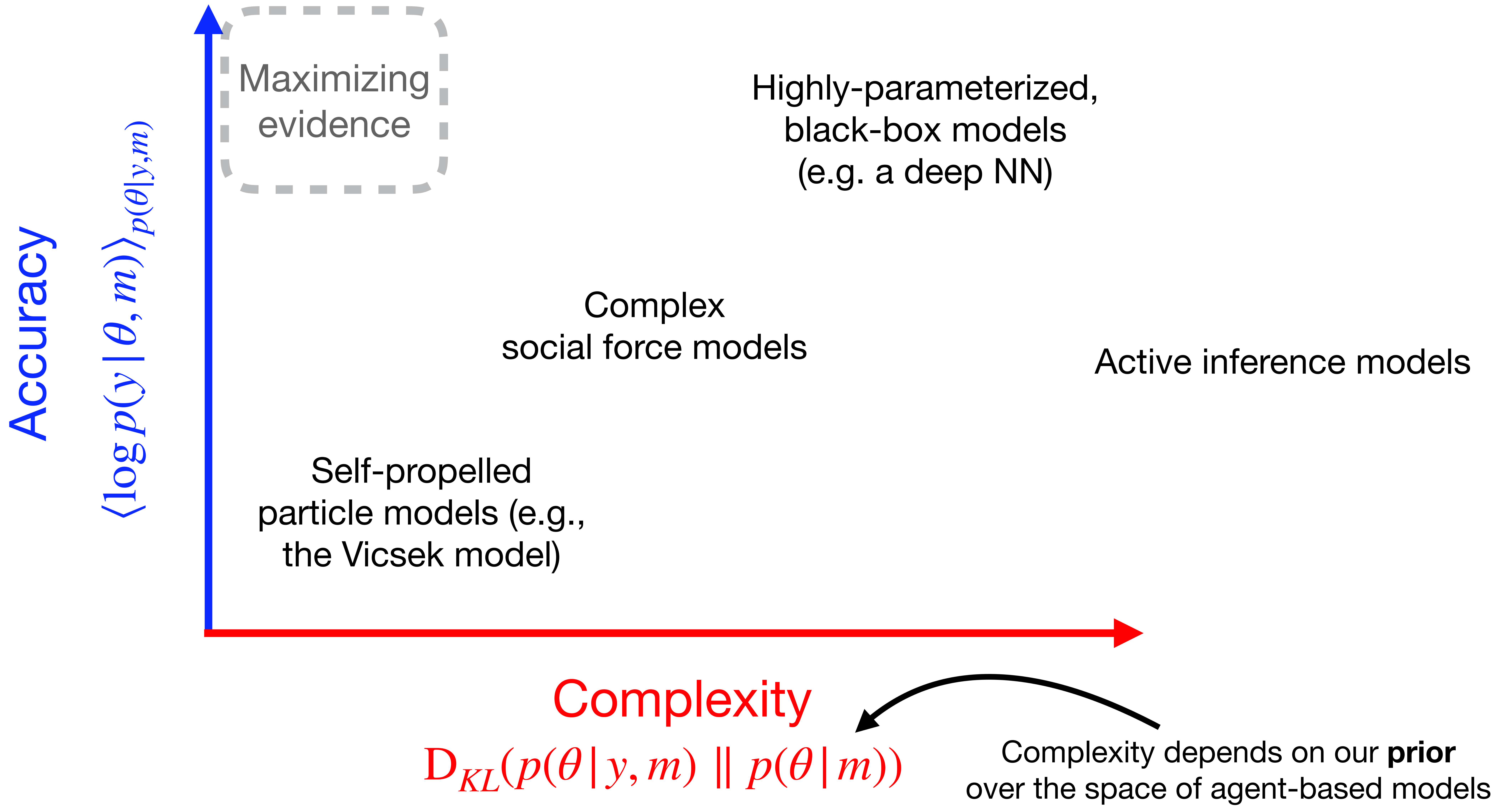
(log) Model evidence = Accuracy − Complexity

How well I fit the data

How many extra bits my explanation need to encode,
relative to my "baseline" expectations about explanations

$$\log p(y \mid m) = \langle \log p(y \mid \theta, m) \rangle_{p(\theta|y,m)} - \mathrm{D}_{KL}(p(\theta \mid y, m) \parallel p(\theta \mid m))$$

The space of agent-based models of collective phenomena



Variational free energy

$$\mathscr{F} = -\int q(\theta) \ln \frac{p(y,\theta)}{q(\theta)}$$

$$\mathscr{F} = \underbrace{\mathrm{D}_{KL}\left(q(\theta) \parallel p(\theta)\right)}_{\text{Complexity}} - \underbrace{\mathbb{E}_{q(\theta)}[\log p(y \mid \theta)]}_{\text{Accuracy}}$$