Bayesian Affect Control Theory of Self

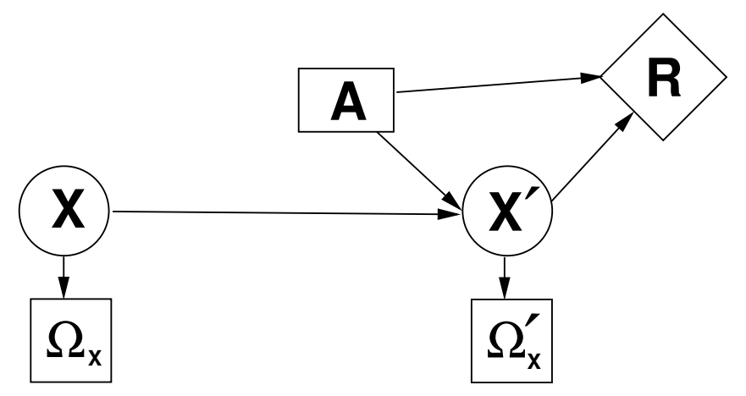
Jesse Hoey[1] and Tobias Schröder[2]

[1]David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Canada
[2]Potsdam University of Applied Sciences, Potsdam, Germany

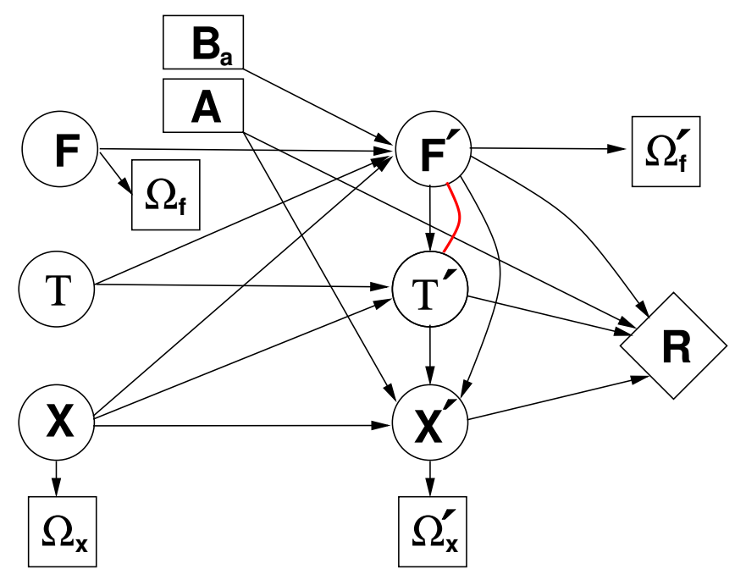
Introduction

- identity and self are key social psychological principles of social interaction and coordination
- important for artificially intelligent agents who:
  - are natural
  - are socially appropriate
  - use subtle human socio-affective skills
- sociological Affect Control Theory of Self **ACT-S** [5]:
  - humans maintain a deep sense of self that:
    - captures emotional, psychological, and socio-cultural sense of being
    - is externalised as a situational identity
    - humans enact identities consistent with their sense of self
    - inauthenticity grows if a person can't enact consistently
- we propose a Bayesian generalization of ACT-S called ***BayesAct-S*** as a foundation for socio-affectively skilled artificial agents, where the self is a probability distribution, allowing an agent to have:
  - multi-modal self: have multiple different identities
  - uncertain self: unsure about who it really is
  - learnable identities: for self and others
  - goal-directed behaviour: based on socio-cultural factors
- we show how ***BayesAct-S*** can underpin artificial agents that are socially intelligent

Partially Observable Markov Decision Process



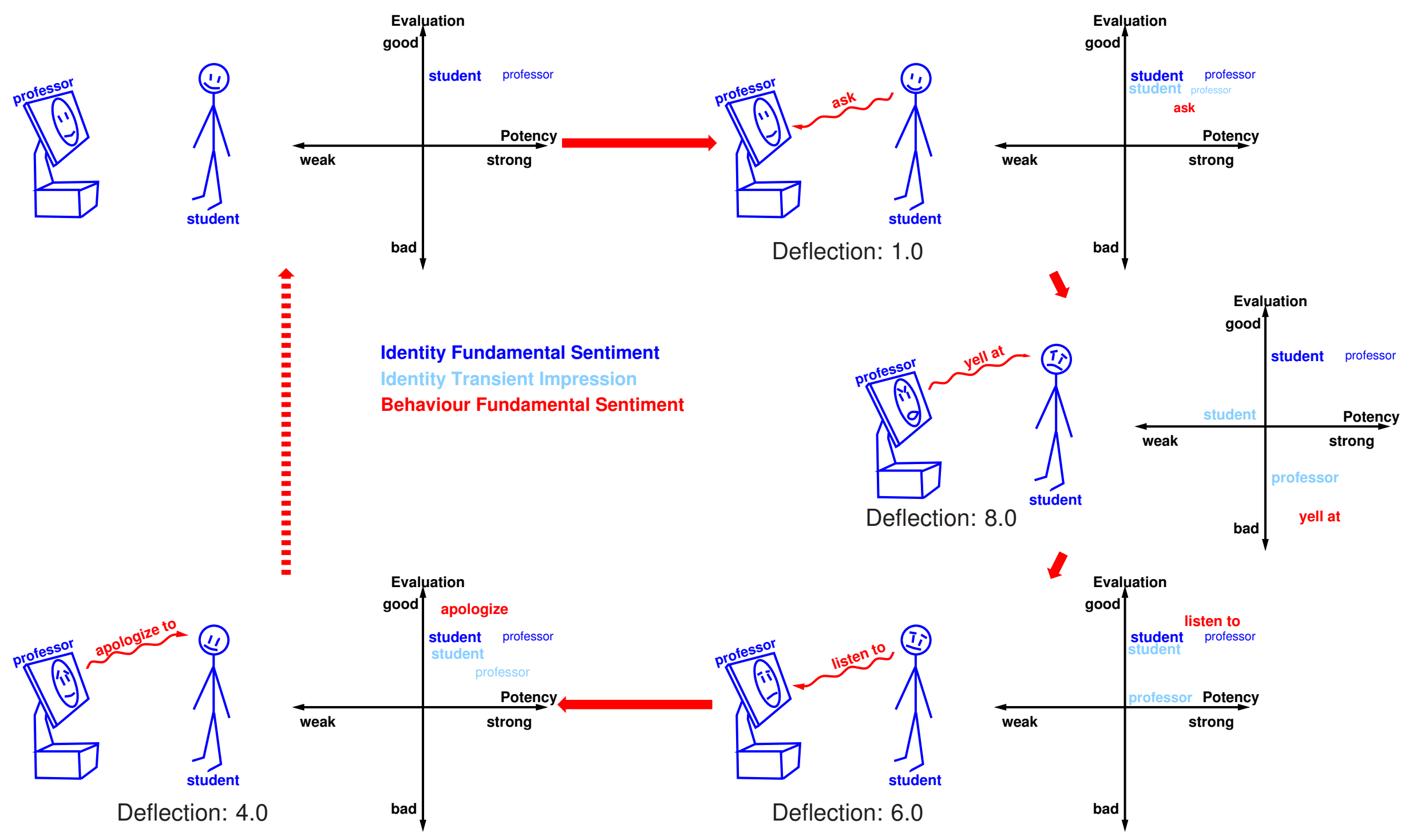
General POMDP



POMDP for *BayesAct*

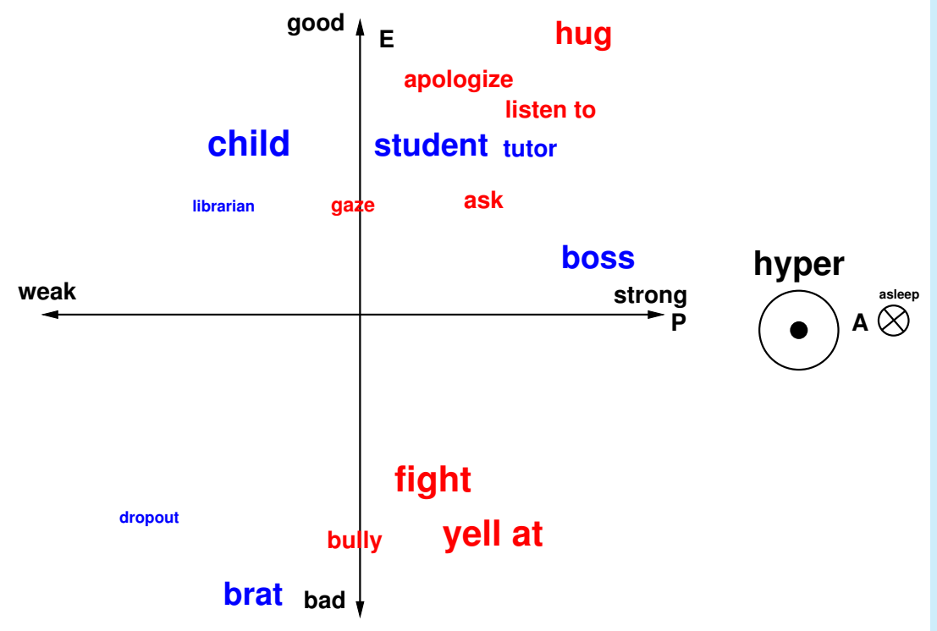
- a policy maps belief states (i.e., distributions over $\mathcal{X}$) into choices of actions, such that the expected discounted sum of rewards is (approximately) maximised
- POMDPs have been used as models for many human-interactive domains (see [3])

Affect Control Theory Example



Sociological Theory

EPA Space [6]



- 3-D **EPA** space [6]
- **E**valuation, **P**otency, **A**ctivity
- shared sentiments across a cultural group
- universal organising principle of human socio-affective experience
  - is compatible with appraisal theories [7]: goal congruence of an event (E), the agent's coping potential (P), and the urgency (A)

Affect Control Theory (ACT) [1]

- Actor-Behaviour-Object (A, B, O) Grammar
- shared fundamental sentiments $(\forall A, B, O)$: $\mathbf{F} \in [-4.3, 4.3]^9$
- transient impressions created by events $A - B - O$ $(\forall A, B, O)$: $\mathbf{T} \in [-4.3, 4.3]^9$
- deflection $D = \sum_i w_i(\mathrm{f}_i - \tau_i)^2$
- prediction $\mathbf{T}_{t+1} = \boldsymbol{M}\mathscr{G}(\mathbf{F}_t, \mathbf{T}_t)$
- $\mathbf{F}, \boldsymbol{M}, \mathscr{G}$: measured empirically [2]

Affect Control Principle: actors work to experience transient impressions that are consistent with their fundamental sentiments

ACT of Self (ACT-S) [5]

- a higher-order level of socio-affective control than ACT
- fundamental self-sentiment ($\mathbf{S_f}$): a person's core (long-lasting) feeling of self
- situational self-sentiment: emphemeral feeling

$$\mathbf{s_s}^T = \sum_{t=0}^{T} w(t, T)\mathbf{f}_a^t$$

composite over recent experiences of self-identity $\mathbf{f}_a$

- accumulated inauthenticity

$$\mathbf{i_a}^T = \sum_{t=0}^{T} w(t, T)\left(\mathbf{f}_a^t - \mathbf{s_f}^t\right) = \mathbf{s_s}^T - \sum_{t=0}^{T} w(t, T)\mathbf{s_f}^t$$

- if $\mathbf{s_f}$ constant and $w(t, T) = \eta^{T-1}$:

$$\mathbf{i_a} = \mathbf{s_s} - \mathbf{s_f}\frac{1}{1-\eta}$$

Affect Control Principle of Self: actors construct situational self-sentiments (by seeking out situations and other actors) to minimize accumulated inauthenticity

Bayesian Generalisation

BayesACT [4]

- fundamental sentiments $\mathbf{F} = \{\mathrm{F}_{ij}\}$ where $\mathrm{F}_{ij}, i \in \{a, b, c\}, j \in \{e, p, a\}$
- transient impressions $\mathbf{T} = \{\mathrm{T}_{ij}\}$
- application states $\mathbf{X}$
- actions: affective ($\mathbf{b_a}$) and cognitive (a)
- transient dynamics $Pr(\boldsymbol{\tau}'|\boldsymbol{\tau}, \mathbf{f}', \mathbf{x}) = \delta(\boldsymbol{\tau}' - \boldsymbol{M}\mathscr{G}(\mathbf{f}', \boldsymbol{\tau}, \mathbf{x}))$
- affect control potential $\varphi(\mathbf{f}', \boldsymbol{\tau}') \propto e^{-(\mathbf{f}'-\boldsymbol{\tau}')^T\Sigma^{-1}(\mathbf{f}'-\boldsymbol{\tau}')}$
- reward function $R(\mathbf{f}, \boldsymbol{\tau}, \mathbf{x}) = R_x(\mathbf{x}) + R_s(\mathbf{f}, \boldsymbol{\tau})$ combines application goals and deflection minimizing goal
- application dynamics $Pr(\mathbf{x}'|\mathbf{x}, \mathbf{f}', \boldsymbol{\tau}', a)$
- observation functions $Pr(\omega_{\boldsymbol{f}}|\mathbf{f}), Pr(\omega_{\boldsymbol{x}}|\mathbf{x})$

generalisation of the affect control principle:

$$\psi(\mathbf{f}', \boldsymbol{\tau}, \mathbf{x}) = (\mathbf{f}' - \boldsymbol{M}(\mathbf{x})\mathscr{G}(\mathbf{f}', \boldsymbol{\tau}, \mathbf{x}))^T\Sigma^{-1}(\mathbf{f}' - \boldsymbol{M}(\mathbf{x})\mathscr{G}(\mathbf{f}', \boldsymbol{\tau}, \mathbf{x}))$$

affective "inertia":

$$\xi(\mathbf{f}', \mathbf{f}, \mathbf{b_a}, \mathbf{x}) \equiv (\mathbf{f}' - \langle\mathbf{f}, \mathbf{b_a}\rangle)^T\Sigma_f^{-1}(\mathbf{x})(\mathbf{f}' - \langle\mathbf{f}, \mathbf{b_a}\rangle)$$

fundamental dynamics:

$$Pr(\mathbf{f}'|\mathbf{f}, \boldsymbol{\tau}, \mathbf{x}, \mathbf{b_a}, \varphi) \propto e^{-\psi(\mathbf{f}', \boldsymbol{\tau}, \mathbf{x}) - \xi(\mathbf{f}', \mathbf{f}, \mathbf{b_a}, \mathbf{x})}$$

BayesACT-S [this paper]

represent $\mathbf{S_s}$ and $\mathbf{S_f}$ as probability distributions

- averaging method (Expressive Order) [1]:

$$\mathbf{s_s}^T = \mathbf{f}_a^T + \eta\sum_{t=0}^{T-1} w(t, T-1)\mathbf{f}_a^t = \mathbf{f}_a^T + \eta\mathbf{s_s}^{T-1}$$

as probability distributions:

$$Pr(\mathbf{s_s}^T) = Pr(\mathbf{f}_a^T) * Pr(\eta\mathbf{s_s}^{T-1})$$

- noisy-Or Method:

$$\mathbf{s_s}' = c\mathbf{s_s} + (1-c)\mathbf{f}_a' \quad \text{where} \quad c \sim Bernoulli(\eta, 1-\eta)$$

as probability distributions:

$$Pr(\mathbf{s_s}') = \int_{\mathbf{s_s}, \mathbf{f}_a'} \sum_{C} Pr(\mathbf{s_s}', c|\mathbf{s_s}, \mathbf{f}_a')Pr(\mathbf{s_s}, \mathbf{f}_a')$$
$$= \eta Pr(\mathbf{s_s}) + (1-\eta)Pr(\mathbf{f}_a)$$

inauthenticity for $\mathbf{s}$:

$$\mathbf{i_a}(\mathbf{s}) = ln\left(\frac{Pr(\mathbf{s_s})}{Pr(\mathbf{s_f})}\right)$$

expected total inauthenticity:

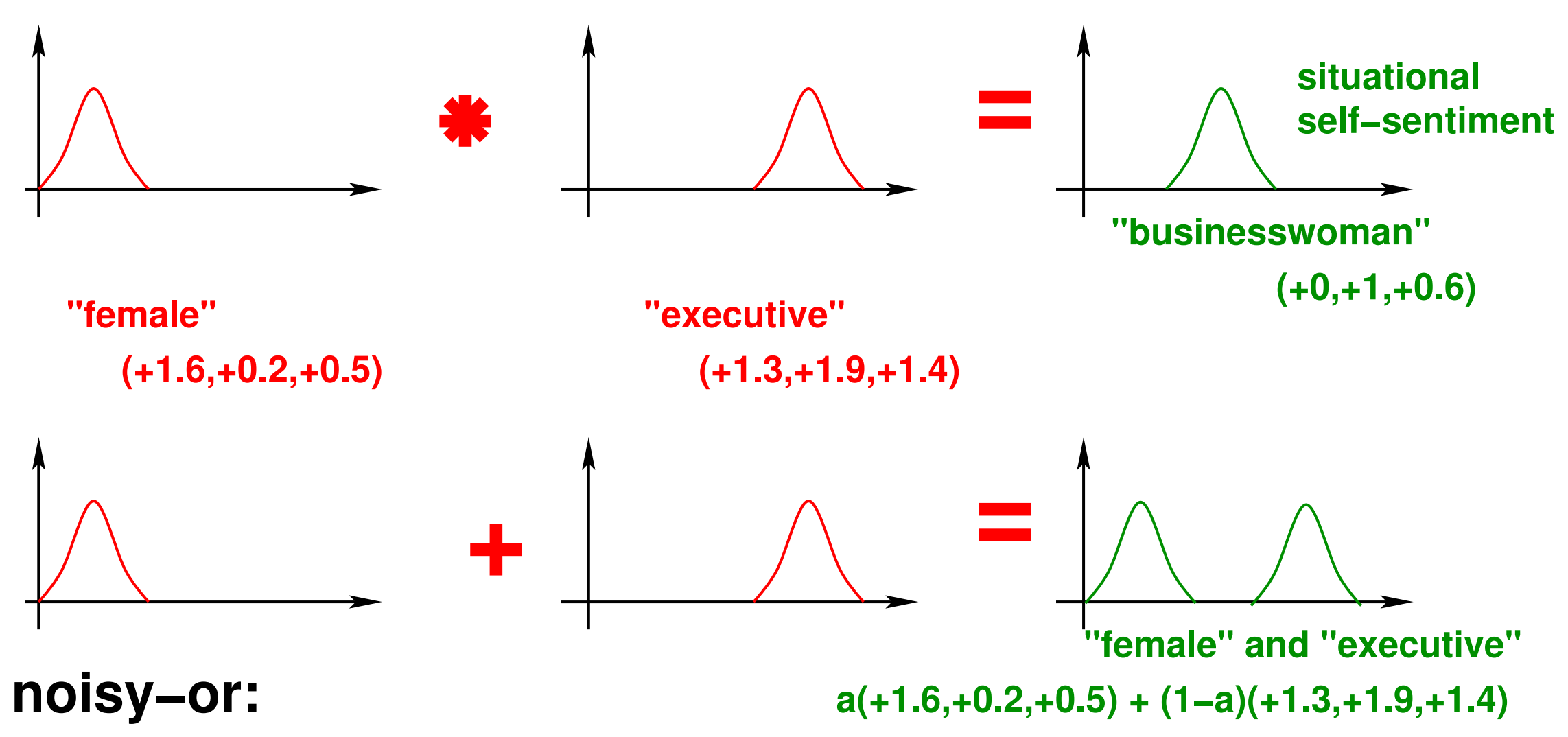
$$\mathbb{E}[\mathbf{i_a}] = \int_{\mathbf{s}} \mathbf{i_a}(\mathbf{s})Pr(\mathbf{s_s})d\mathbf{s}$$

→ Kullback-Leibler (KL) divergence between $\mathbf{s_f}$ and $\mathbf{s_s}$

BayesAct-S selects interactions that will minimize the expected inauthenticity, $\mathbb{E}[\mathbf{i_a}]$

Averaging vs. Noisy-OR

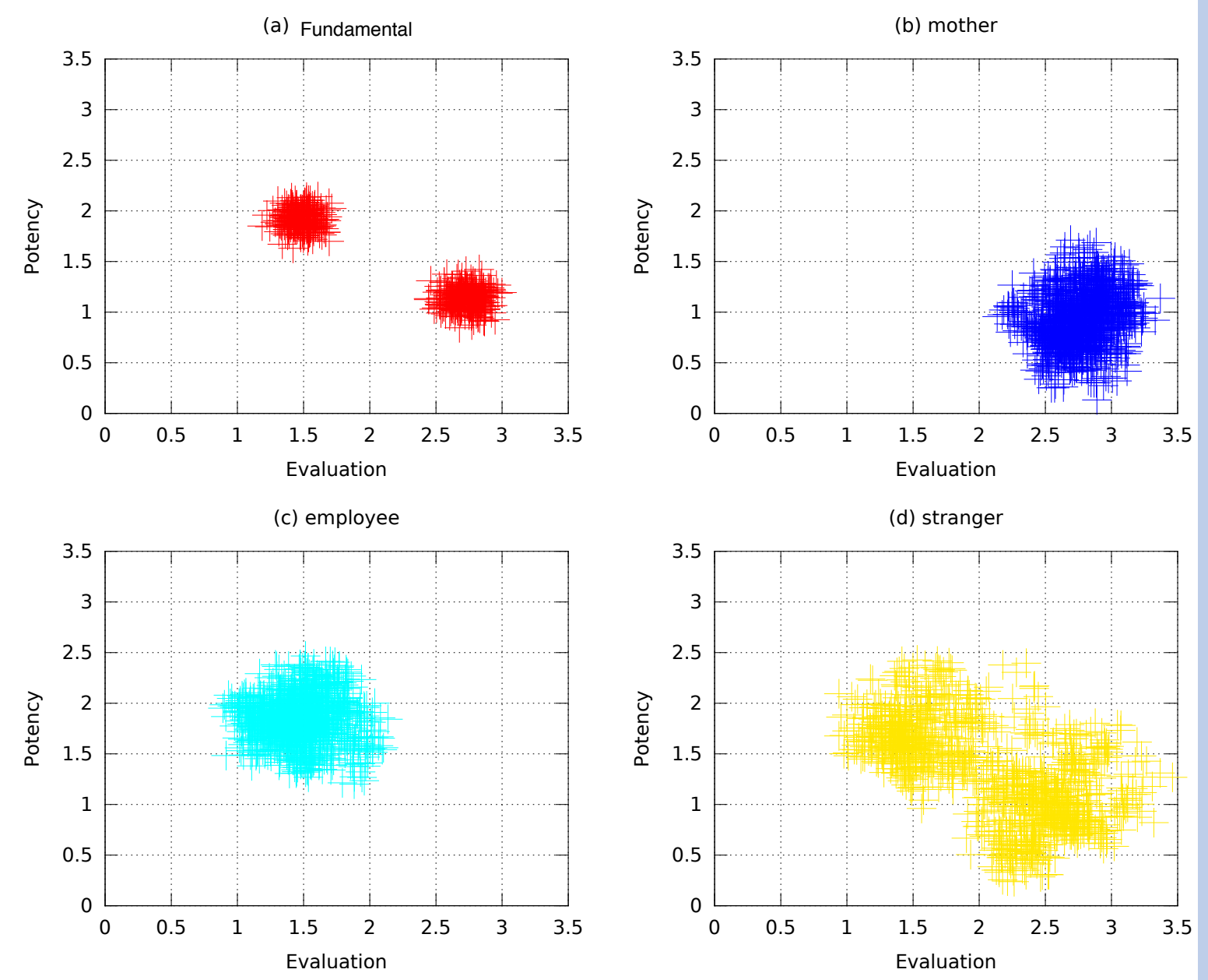
two methods for computing situational self-sentiments:



Simulations

- a female agent with a mixture of 2 identities
  - *daughter* (EPA=[2.73, 1.13, 1.28])
  - *employer* (EPA=[1.48, 1.93, 0.74])
- two client identities
  - *mother* (EPA=[3.12, 2.98, 1.44]
  - *employee* (EPA=[1.88, 0.05, 0.84]).

after 20 interactions ↗ agent's situational self sentiment changes based on the other agent



(a) fundamental (mix between female and employer)
(b) mother → *female/employer* feels like daughter
(c) employee → *female/employer* feels like employer
(d) stranger → *female/employer* feels both

KL-divergences ⟶ shows **who** the agent will interact with next

agent recently interacted with ↓	will interact with next:		
	employee	stranger	mother
employee	3.09	2.46	**2.17**
stranger	2.37	2.96	**2.27**
mother	**2.18**	2.38	2.80

Conclusion

the socio-affective agent model ***BayesAct-S*** :

- is used for fast, heuristic, learnable agent interaction
- is how to "get along" with other agents in a social world
- unifies the cognitive (individual) and affective (social)
- gives agents a societal guide for selecting goals, settings, institutions and individuals to interact with

References

[1] David R. Heise. *Expressive Order: Confirming Sentiments in Social Actions*. Springer, 2007.

[2] David R. Heise. *Surveying Cultures: Discovering Shared Conceptions and Sentiments*. Wiley, 2010.

[3] Jesse Hoey, *et al.* People, sensors, decisions: Customizable and adaptive technologies for assistance in healthcare. *ACM Trans. Interact. Intell. Syst.*, 2(4):20:1–20:36, January 2012.

[4] Jesse Hoey, Tobias Schröder, and Areej Alhothali. Bayesian affect control theory. In *Proc. of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII 2013)*, 2013.

[5] Neil J. MacKinnon and David R. Heise. *Self, identity and social institutions*. Palgrave and Macmillan, New York, NY, 2010.

[6] Charles E. Osgood, William H. May, and Murray S. Miron. *Cross-Cultural Universals of Affective Meaning*. University of Illinois Press, 1975.

[7] Kimberly B. Rogers, Tobias Schröder, and Christian von Scheve. Dissecting the sociality of emotion: A multi-level approach. *Emotion Review*, 6(2):124–133, 2014.

David R. Cheriton School of Computer Science, University of Waterloo

email: jhoey@cs.uwaterloo.ca
WWW: bayesact.ca