AI Horizons Institute

Neuro- and Developmental- Inspired AI White Paper

(Adapted from the preliminary proposal for the STRONG AI NSF Institute)

Limitations of Today's GenAI Systems. Existing LLMs based on autoregressive transformers (e.g., GPT models, PaLM, PaLM-E, etc.), must output tokens to "think" in a sequence of forward passes. This results in poor planning and being highly sensitive to the prompt [1]. For example, telling an LLM to "think step-by-step" often produces better results [2] because the model spends more time "talking" as the generated next token is concatenated onto the input prompt. In a sense, their only ability to think is "aloud" which impairs instructability. Alignment is enforced via post-hoc ?ne-tuning [3, 4] rather than being a fundamental component from the start. Systems cannot be easily updated without catastrophically forgetting past knowledge, resulting in an inability to learn about recent events [5]. While retrieval augmented generation (RAG) methods enable pre-trained LLMs to access an external vector database to access new knowledge [6], this is a crude approximation to what is needed, where data would regularly be consolidated from a growing database into the model for long-term storage. Likewise, multi-modal information is injected into LLMs in a post-hoc manner, which we hypothesize impairs their ability to learn grounded concepts. Systems are highly sensitive to training data bias, often amplifying them [7]. LLMs cannot assess when they should engage in creative prose rather than fact-based processing, where they will invent plausible-sounding justifications with no basis in fact. While this is often called "hallucination," in humans, this is known as *confabulation* and is caused by multiple forms of brain damage. Today's GenAI systems depend on data generated by people – they cannot generate knowledge themselves. They are incapable of self-correcting errors in their reasoning without external feedback [8]. Rather than blindly ingesting data, GenAI systems need to be motivated, curious lifelong learners that intrinsically facilitate alignment with human values. We propose to address these limitations by taking inspiration from the functional neural modules and principles that enable alignment, instructability, and grounding in the human brain. These suggest we need to incorporate architectural inductive biases into networks to enable these abilities, where they are learned throughout the model's life rather than as a post-hoc band-aid applied to pre-trained models.

Endowing LLMs with Self-Talk, Introspection, & Metacognition. In humans, the OFC is essential for the maintenance of normative alignment. OFC damage causes compulsive behavior, an inability to plan, confabulation [9], and a loss of sympathy. We propose to create an OFC analog, which acts to regulate the network to inhibit behaviors not aligned with established norms. Just like in humans, the artificial OFC will require the ability to incorporate new rules of safety without giving up other safety rules. OFC also plays an essential role in uncertainty estimation and metacognition, enabling people to "know what they don't know" [10]. Endowing the model to generate self-talk and to "think" in concept embeddings, rather than in sampled tokens, and then having the model assess these "thoughts" to determine which should be emitted has the potential to greatly reduce prompt engineering and enable auditing these thoughts to prevent undesirable behaviors. We want to explore the possibility to create new non-transformer architectures that integrate an artificial OFC,

thereby building alignment directly into the network as a form of inductive bias, rather than the existing post-hoc methods used today such as instruction fine-tuning and RLHF [3]. The artificial OFC will greatly reduce confabulation by providing metacognition.

Lifelong Learning of Episodic and Semantic Memories. An essential property of general intelligence is the ability to continuously learn, which today's LLMs lack. Fine-tuning LLMs on new data leads to catastrophic forgetting of past abilities [5]. The hack used to overcome this limitation are RAG models, where a vector store acts as a crude episodic memory for retrieving information to be injected into the prompt, with the LLM never updated from the vector store. We argue that an explicit episodic memory is needed from the start of learning, where data should be consolidated from an episodic memory to form the model's semantic knowledge via learning. In our model, semantic memory is derived from episodic memory in a self-supervised manner, where similar episodic memories facilitate tying past experiences into semantic concepts. For example, the episodic memory would cache the individual experiences with a specific object category, e.g., a strawberry, and then these are grouped over time in a self-supervised manner to obtain semantic representations. This consolidation will occur during sleep phases, which is when episodic memories stored in the hippocampus are transferred and consolidated into cortical regions in mammals [11]. We hypothesize that this architecture will greatly outperform existing RAG-based approaches, which are now widely employed by industry. This thrust builds upon our team's extensive work in continual learning inspired by sleep and memory consolidation mechanisms to enable progressive knowledge accumulation without catastrophic forgetting.

Teaching AI to be Socially Aware. Socially aware agents understand how they impact others and how they impact them. It is a critical component of empathy. Primates are social animals taught the norms of their societies by their families. Monkeys raised in isolation are unable to interact with others and will mutilate their own offspring [12, 13]. Similarly, human orphans raised in isolation have symptoms of autism spectrum disorder and disinhibited social engagement as adults at much higher rates than controls despite later being adopted [14]. While our artificial OFC is necessary for overcoming many of GenAI's problems, these findings suggest architecture alone will not suffice. A GenAI needs to be raised in a social environment for it to acquire appropriate skills and alignment. While there are extreme limitations to today's LLMs that achieve alignment via post-hoc mechanisms, we want to explore using them as surrogate parents to teach our models to be inquisitive and aligned during their developmental phase. During "wake" phases, the two models will engage in multi-modal dialogues, where the teacher will present multi-modal inputs to our model, which will cache these experiences in episodic memory. To make this approach efficient, we could employ methods that enable the network to convert between a recurrent processing mode for generation (wake phase) and a parallel offline mode for training (sleep phase) [15]. This work will be heavily informed by developmental neuroscience. Supporting this idea, it was recently shown that GPT-4 can be used to create curated datasets for efficient training of other LLMs [16], albeit without alignment. We also want to study directly learning from humans once our model has acquired communication skills.

Grounding via Intrinsically Multi-Modal AI Systems. To avoid the infinite regression problem,

mathematics has learned over the past 200 years that inherently primitive objects should be undefined, where they gain their power as abstractions via their relationships [17]. Similarly, the brain is thought to achieve grounding by having the senses predict each other, i.e., the relationships between representations from multiple senses enable grounding. This approach has been used to great success in deep learning, where language and vision encoders are trained to produce the same embeddings [18]. Brain regions responsible for each sense also predict future observations for each modality, which is used for self-supervised learning in both language and vision. We propose to explore a novel neural network architecture based on this idea, where all modalities are trained jointly, but the architecture itself has functional segregation in a manner similar to how the brain has visual, linguistic, auditory, and multi-modal integration regions. Each functional module is trained to predict future unimodal observations and the outputs of other modalities, e.g., the embeddings for the word "dog", the sound of barking, and a photo of a dog would all have similar embeddings. While this has been done for representation learning in non-GenAI models that are later frozen and plugged into GenAI systems, in our model, this serves as an internal mechanism for grounding, where everything is trained jointly. Rather than a single loss, there will be many for each uni-modal, multi-modal, and generative functional modules.

Learning in Today's LLMs is Flawed. Today's GenAI systems depend on existing data generated by people and on what is fed to them – they cannot ask for information, let alone generate knowledge or evaluate the information that they receive or produce. Today's LLMs are trained in three phases. First, they are trained in a self-supervised manner using next-token prediction, where tokens represent words or sub-words, typically from a scrape of the entire Internet. The data is typically curated only to reduce redundancy, but there is often a great deal of "toxic" text incorporated and the LLMs have no mechanism to verify the quality of the data or the veracity of the information. After this phase, the LLM can complete documents but directly interact with others (i.e., be an assistant). The second phase turns it into an assistant using human-produced prompt-response dialogues, which are expensive to produce. Lastly, an additional "alignment" step is done to encourage the model to emit responses consistent with their creator's preferences, with the most popular method being reinforcement learning from human feedback (RLHF), which requires additional data from humans for ranking output responses. While RLHF has given us powerful LLMs (e.g., ChatGPT), it is a post-hoc band-aid. Systems cannot be easily updated with new knowledge, alignment is not done throughout, systems can be "jail-broken," and systems are not trained to have metacognitive awareness [19], resulting in hallucinations. Most importantly, current systems do not know what they do not know, nor how to ask questions to learn new information or to clarify instructions given by their users. They cannot self-correct errors in their reasoning without external feedback. They also lack social awareness. Socially aware agents understand how they impact others and how others impact them [20]. It is a critical component of both grounding and alignment (i.e., empathy). To address these issues, we propose a new way of training GenAIs inspired by learning in children.

How Human Children Become Grounded and Aligned. Although having the necessary neural architecture, e.g., the PFC, is necessary for humans to become *grounded* and *aligned*, it does not suffice. The environment in which a human child is raised plays an equally critical role. Humans are social animals taught both the knowledge and the norms of their societies by parents, teachers, and even peers. Indeed, humans and other primates raised in isolation show severe cognitive and social impairments [21-23]. More than this, there is much that children themselves bring to the learning table: Developmental research suggests that children come into the world highly motivated to learn

both conceptual information (*grounding*;[24] and the norms of their society (*alignment;* [24]). In this sense, children are naturally *instructible* (i.e., prepared to learn; [26]). Indeed, even human infants are predisposed to imitate others [27], to attend to patterns [28—30], to track motion [31-32], and to show special attention to social information (e.g., human faces [33]; language, 34]; and emotions [35]). As children develop, they also become intensely inquisitive beings, exploring their environments, seeking out interactions, and asking questions to fill in knowledge gaps [26; 36-37]. They in turn become increasingly adept at evaluating the information they receive and recognizing the extent and limits of their own knowledge and skills; that is, what is known and what needs to be known (i.e., metacognitive awareness [19]). They also become increasingly aligned with the norms and expectations of their local environments. Current AI systems lack this inquisitive nature and this metacognitive and social awareness, all of which are paramount for becoming *grounded* and *aligned*. The outline provided as an Appendix further elaborates on what we can learn from Child Development to develop the next generation of AI.

References

[1] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. Ribeiro, and Y. Zhang. "Sparks of arti□cial general intelligence: Early experiments with GPT-4." In: *arXiv preprint arXiv:2303.12712* (2023).

[2] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, and D. Zhou. "Chain-of-thought prompting elicits reasoning in large language models." In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 24824–24837.

[3] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama,
A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. "Training language models to follow instructions with human feedback." In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 27730–27744.

[4] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. E. Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hat□eld-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, and J. Kaplan. "Constitutional AI: Harmlessness from AI feedback." In: *arXiv preprint arXiv:2212.08073* (2022).

[5] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, et al. "Palm-e: An embodied multimodal language model." In: *arXiv preprint arXiv:2303.03378*
(2023).

[6] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. "Retrieval-augmented generation for knowledge-intensive NLP tasks." In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 9459–9474.

[7] M. Hall, L. van der Maaten, L. Gustafson, M. Jones, and A. Adcock. “A systematic study of bias ampli□cation.” In: *arXiv preprint arXiv:2201.11706* (2022).

[8] J. Huang, X. Chen, S. Mishra, H. S. Zheng, A. W. Yu, X. Song, and D. Zhou. “Large Language Models Cannot Self-Correct Reasoning Yet.” In: *arXiv preprint arXiv:2310.01798* (2023).

[9] A. Schnider. “Spontaneous confabulation and the adaptation of thought to ongoing reality.” In: *Nature Reviews Neuroscience* 4.8 (2003), pp. 662–671.

[10] D. R. Bach and R. J. Dolan. “Knowing how much you don’t know: a neural organization of uncertainty estimates.” In: *Nature reviews neuroscience* 13.8 (2012), pp. 572–586.

[11] T. L. Hayes, G. P. Krishnan, M. Bazhenov, H. T. Siegelmann, T. J. Sejnowski, and C. Kanan. “Replay in deep learning: Current approaches and missing biological elements.” In: *Neural computation* 33.11 (2021), pp. 2908–2950.

[12] H. F. Harlow, R. O. Dodsworth, and M. K. Harlow. “Total social isolation in monkeys.” In: *Proceedings of the National Academy of Sciences* 54.1 (1965), pp. 90–97.

[13] D. Blum. *The monkey wars*. Oxford University Press, 1995.

[14] E. J. Sonuga-Barke, M. Kennedy, R. Kumsta, N. Knights, D. Golm, M. Rutter, B. Maughan, W. Schlotz, and J. Kreppner. “Child-to-adult neurodevelopmental and mental health trajectories after early life deprivation: the young adult follow-up of the longitudinal English and Romanian Adoptees study.” In: *The Lancet* 389.10078 (2017), pp. 1539–1548.

[15] Y. Sun, L. Dong, S. Huang, S. Ma, Y. Xia, J. Xue, J. Wang, and F. Wei. “Retentive network: A successor to transformer for large language models.” In: *arXiv preprint arXiv:2307.08621* (2023).

[16] Y. Li, S. Bubeck, R. Eldan, A. Del Giorno, S. Gunasekar, and Y. T. Lee. “Textbooks Are All You Need II: phi-1.5 technical report.” In: *arXiv preprint arXiv:2309.05463* (2023).

[17] A. Tarski and J. Tarski. *Introduction to Logic and to the Methodology of the Deductive Sciences*. 24. Oxford University Press, USA, 1994.

[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. “Learning transferable visual models from natural language supervision.” In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763.

[19] J. Flavell. *Cognitive Development*. Englewood Cliffs, NJ: Prentice Hall (1977)

[20] M. Tomasello. Joint attention as social cognition. In C. Moore & P. J. Dunham (Eds.), *Joint attention: Its origins and role in development* (pp. 103–130). Lawrence Erlbaum Associate, (1995).

[21] H. F. Harlow. The nature of love. *American Psychologist, 13*(12), 673–685. https://doi.org/10.1037/h0047884, (1958)

[22] C. A. Nelson et al., The Neurobiological toll of early human deprivation. In the Monographs of the Society for Research in Child Development, 76(4), pp. 127-146 (2011). https://doi.org/10.1111/j.1540-5834.2011.00630.x

[23] M. Rutter et al., Deprivation-specific psychological patterns: Effects of institutional deprivation. In the Monographs of the Society for Research in Child Development, 75(2). (2010).

[24] H. Wellman & S. A. Gelman. Knowledge acquisition in foundational domains. In W. Damon (Ed.), *Handbook of child psychology: Vol. 2. Cognition, perception, and language* (pp. 523–573). John Wiley & Sons, Inc. (1998)

[25] L. Schmidt et al., Young people's more permissive views about marijuana: Local impact of state laws or national trend? American Journal of Public Health, 106(8), 1498-503 (2016) doi: 10.2105/AJPH.2016.303153.

[26] J. Piaget. *The construction of reality in the child.* (M. Cook, Trans.). Basic Books. (1954) https://doi.org/10.1037/11168-000

[27] A. Melzoff & M. K. Moore. Imitation of facial and manual gestures by human neonates. Science, 198(4312); 74-8. (1977) doi: 10.1126/science.897687.

[28] M. L. Courage et al., Infants' attention to patterned stimuli:Developmental change from 3 to 12 months of age. Child Development, 77(3): 680-95. (2006) DOI:10.1111/j.1467-8624.2006.00897.x

[29] R. L. Fantz, The origin of form perception. *Scientific American, 204*(5), 66–72. (1961). https://doi.org/10.1038/scientificamerican0561-66

[30] F. Xu & V. Garcia. Intuitive statistics by 8-month-old infants. Proc Natl Acad Sci U S A. (2008). Apr 1;105(13):5012-5. doi: 10.1073/pnas.0704450105. Epub 2008 Mar 31. PMID: 18378901; PMCID: PMC2278207.

[31] Kellman, P. J., & Spelke, E. S. (1983). Perception of partly occluded objects in infancy. *Cognitive Psychology, 15*(4), 483–524. https://doi.org/10.1016/0010-0285(83)90017-8

[32] Johnson, S. P., & Aslin, R. N. (1995). Perception of object unity in 2-month-old infants. *Developmental Psychology, 31*(5), 739–745. https://doi.org/10.1037/0012-1649.31.5.739

[33] Farroni T, Johnson MH, Menon E, Zulian L, Faraguna D, Csibra G. Newborns' preference for face-relevant stimuli: effects of contrast polarity. Proc Natl Acad Sci U S A. 2005 Nov 22;102(47):17245-50. doi: 10.1073/pnas.0502205102. Epub 2005 Nov 11. PMID: 16284255; PMCID: PMC1287965.

[34] Vouloumanos A, Werker JF. Listening to language at birth: evidence for a bias for speech in neonates. Dev Sci. 2007 Mar;10(2):159-64. doi: 10.1111/j.1467-7687.2007.00549.x. PMID: 17286838

[35] Heck A, Hock A, White H, Jubran R, Bhatt RS. Further evidence of early development of attention to dynamic facial emotions: Reply to Grossmann and Jessen. J Exp Child Psychol. 2017 Jan;153:155-162. doi: 10.1016/j.jecp.2016.08.006. Epub 2016 Sep 27. PMID: 27686256; PMCID: PMC5191505.

[36] Chouinard, M. M. (2007). Children's questions: A mechanism for cognitive development: III. Diary study of children's questions. *Monographs of the Society for Research in Child Development, 72*(1), 45–57. https://doi.org/10.1111/j.1540-5834.2007.00415.x

[37] Menendez D, Klapper RE, Golden MZ, Mandel AR, Nicholas KA, Schapfel MH, et al. (2021) "When will it be over?" U.S. children's questions and parents' responses about the COVID-19 pandemic. PLoS ONE 16(8): e0256692. https://doi.org/10.1371/journal.pone.0256692

APPENDIX

What Can AI Learn from Child Development?

(by Karl S. Rosengren; Isobel Heck; Christopher Kanan)

Context matters!

- Humans are shaped by the context they are in (past / present / future)
- While there are some universals – these are influenced by context / experience
- Child is influenced by different levels (Bronfenbrenner, 1979)
- Parents/ Peers etc.,
- Interaction of parents/peers etc.
- Extended family, neighbors, social class, etc.,
- Cultural attitudes and beliefs
- The nature and degree of influence on the child of these levels varies over the life course

Humans are complex systems

- Need to consider systems separately and how they interact
- There is a lot of variability over the course of development
- Across children of the same age
- In the overall pattern of change

Systems emerge and develop at different rates, and are influenced by other systems and contexts to various degrees

Different perceptual systems pick-up overlapping (but not-completely redundant information) –
Provides convergence / flexibility in learning

Human infants are not blank slates

- Over course of evolution infants come in to the world with:
- Particular skills
- Particular biases / tendencies

Infants are active, fast, and highly flexible learners

- Desire to learn / search for causal relations

How do we know we know what infants and children know?

Can't ask directly
Must infer from indirect means
Advances in the study of infant cognition
Infants have built in preferences (Preference Studies)
Infants get bored with repeated stimulation (Habituation Studies)
Infants presented with a repeating event
Looking time
fMRI

Infants can interpret a series of events (Violation of Expectation Studies

- Similar to habituation approach, but show actions with possible / impossible outcomes
- (Outcomes consistent / inconsistent with adult expectations)

- **Longer looking time** – indicative of VOE (surprise)
- **Pupillometry** – Larger pupil dilation for VOE
- **fMRI** – change in brain activity

What have researchers learned from these techniques

- **Infants Prefer:**
- **High contrast** visual images vs Low contrast
- Things that **move** vs things that don't
- **Novel** over familiar (most of the time)
- **Faces** over nonfaces (or inverted or scrambled faces)
- **Human voices** over other sounds
- **Biological motion** over non-biological motion
- **Match** vs Mismatch of Visual & Auditory Information (intermodal)
- **Senses work together from early on**

What comes built in to the infant?

- **Constraints** – Factors that facilitate or limit aspects of development
- **Maturation** – can't be accelerated significantly
  Brain development -
- Physical development – different systems develop at different rates
- **Body** –
- Physical structures (skeletal/muscular) influence "end state" of certain systems
- Influence the actions (and related cognitions) that are possible (Embodiment)
- **Certain Preferences**

- **Drive to learn**
- Statistical learning
- Trial and Error Learning
- Question asking
- Increased automaticity of cognitive processes
- Acquisition of more complex strategies
- Social Cultural Transmission
-
- **Core Knowledge**
- **Characteristics of Systems**
- **Innate** / Built in
- Present in **infancy** and across the **life span**
- Can be found in individuals **across cultures**
- Can be found in **non-human species**
- Serve as **building blocks** for more advanced cognition

6 Systems:
1 – Ability to parse **objects** and their interactions
2- Attention to **shape** / form
3 – Distinguishing **animate** from inanimate (Agency)
4 – Attention to **social** partners
5 – Attention to locations / **space**
6 – Attention to **numbers**

- **Cognitive Biases**
  - Search for causality – (Baillargeon / Spelke / Gelman)
  - Automatically assume causal link between events that occur close in proximity
    - Magical Beliefs / superstition
    - Essentialism (Gelman)
  - Belief that things have an underlying **essence** that determines their nature
    - Stereotypes (Gelman) / Contagion (Nemeroff & Rozin)
  - Teleology (Keleman)
- Assume things have a purpose / made by design
  - Animism (Piaget / Guthrie)
    Assumption things are animate

- What Develops?

- Basic Cognitive Functions
- Naïve / Intuitive Theories (Wellman & Gelman, 1998)
- Based on causal model / Distinct Domain
- Theory of Mind (TOM)
- Understanding of others' thoughts / actions / intentions
-
- Executive Functioning (control of cognitive processes)
- Metacognition
- Knowing what you know and what you don't know

- Social Cognition

  The process of socialization refers to the process that leads individuals in a society to behave in ways that conform to the society's norms, values, attitudes and beliefs. Socialization agents provide both direct and indirect influences. Direct influences include social referencing, modeling, and providing an environment where emotions are expressed and reacted to on a regular basis. Social referencing refers to situations where a child (or other individual) looks to a caregiver after some experience. For example, a child may go down a slide too fast and land on their bottom at the end of slide. In this situation, the child may look to their caregiver (social reference) to see if the caregiver is concerned or not. If the caregiver expresses worry the child may cry, if the caregiver smiles at the child, the child may be less likely to cry. A child may be simultaneously exposed to a variety of different social cues at any given time and must integrate and determine how to use these cues. For example, imagine the child goes down the slide in front of both a mother and a father – and the parents provide different responses to the child bottoming out. Should they weight one parents' response more?

  Parent-child conversations also play an important role in the defining of different social groups (e,g, race, ethnicity, social class) and transmitting of cultural values and ideologies. Children from a young age begin to recognize, represent, and reason about group based patterns of power, status, and wealth (Heck, Shutts, & Kinzler, 2022)

  Emotional Expression / Understanding
  The way that children come to experience and interpret emotions is greatly influenced by the environment that the child is in and the interactions they have with people in that environment

- Research in this area, not only focuses on the child's expression and understanding of emotion, but how people (caregivers, teachers, peers, society at large) serve as socialization agents to model and or teach about emotion

Understanding Others is a component of **Social Cognition** —
All the ordinary ways in which we make sense of the behavior of other people.

- Includes understanding of *individual psychology*: each of us has our own subjective thoughts, feelings, perspectives, goals, emotions, identities, etc.

Our social cognition is important for:

- Perspective taking
- Empathy
- Morality
- Communication
- Collaboration
- Relationships

How do we learn from others?

Foundations:
joint attention, gaze following (by 9 months perhaps much earlier)
Imitation
Newborns imitate facial gestures
Imitation of actions common in older children
Learning norms
Children can infer a norm from a single action!
Learning other social information

Pons, Harris, and De Rosnay (2004)

- **Recognition** (by 3 to 4 years): Recognize and name emotions based on facial expressions and other cues
- **External cause** (by 3 to 4 years): Understand how external causes in the world shape emotions
- **Desire** (by 3 to 5 years): Appreciate that people's emotion states depend on their desires; different people can have different emotions if they have different desires.
- **Belief** (by 4 to 6 years): Understand that people's emotions are based on their beliefs about the world (whether true or false)
- **Reminder** (3 to 6 years): Understand relation between memory and emotion. Elements of a present situation can reactivate emotions tied to a past situation.
- **Regulation** (changes over early childhood): Younger children tend to refer to behavioral strategies; older children start acknowledging that psychological strategies (distraction, breathing, reframing) can be more effective.
- **Hidden emotions** (4 to 6 years): Outward expressions of emotion may not always match what someone is feeling inside.
- **Mixed emotions** (around 8 years): Understand that people can experience multiple and even conflicting emotions simultaneously.
- **Morality** (around 8 years): Understand that morally reprehensible actions can lead to negative feelings and that morally praiseworthy actions can lead to positive feelings.
-

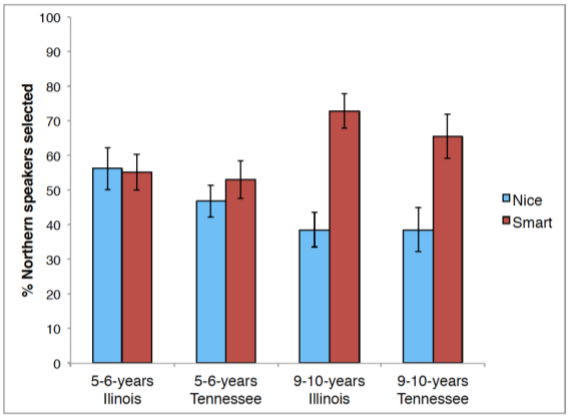
- Knowledge of groups / norms /
- Moral vs Conventional issues

Social groups and identities within these groups play an important role in socialization
Individuals identifying with a group provide the foundation for:
Emergence of ingroup vs outgroup
Thinking, attitudes, biases, stereotypes

Language can highlight the value of groups in terms of
Gender, race, age
Even young infants can distinguish these categories
people see language as *fundamentally* differentiating groups of people
language discrimination can become routinized

Plays a role in development of trust
native-accented speech often viewed as more credible
legal system has been found to devalue testimony of "non-standard" speakers

Kinzler & DeJesus (2013) - Accent attitudes



Children introduced to American English speakers with Northern vs. Southern accents

Children also use language and accent to make predictions about status, leadership, and nationality

Implications:

- What can AI learn from Development?

Are there ways to build in "core knowledge" / "constraints"/ "biases" that make AI smarter?

Are there ways to build in connections between systems?
- Vision / written language / spoken language

Are there ways to build in Metacognition?
- And teach AI systems to ask questions to fill in gaps in knowledge?

Are there ways to build in contextual information/influences?

- Simulating the role of parents / cultural norms and practices

Conclusions:

Developmental research suggests that children come into the world highly motivated to learn both conceptual information (*grounding*) and the norms of their society (*alignment*).
As children develop, they become intensely inquisitive beings, exploring their environments, seeking out interactions, and asking questions to fill in knowledge gaps. They in turn become increasingly adept at evaluating the information they receive and recognizing the extent and limits of their own knowledge and skills; that is, what is known and what needs to be known (i.e., metacognitive awareness). Current AI systems lack this inquisitive nature and this metacognitive awareness, both of which are paramount for becoming *grounded* and *aligned*

Selected References

Fantz, R.L. (1961). The origin of form perception. *Scientific American*, *204*(5) 146, 66-73.

Kourtzi Z., Kanwisher N. (2001). Representation of perceived object shape by the human lateral occipital complex. Science 293, 1506–1509 10.1126/science.1061133

J. Piaget. *The construction of reality in the child*. (M. Cook, Trans.). Basic Books. (1954)

Saffran et al (2007) Dog is a dog: Infant rule learning is not specific to language, Cognition, 105, 669-680.

Spelke, E. (2017). Artificial intelligence and human minds: Perspectives from Young Children.

Spelke, E. & Kinzler, K. (2007). Core knowledge. *Developmental Science, 10,* 89-96.