Living Alignment

Zeb Kurth-Nelson[*1], Steve Sullivan[*2], Nenad Tomašev[3], Erman Misirlisoy[4],
Joel Leibo[3], Matthew Nour[5], and Marc Guitart-Masip[6]

[1]Prefrontal.ai, London, UK
[2]Oregon Health and Science University, Portland, OR
[3]Google DeepMind, London, UK
[4]Independent
[5]Microsoft AI, London, UK
[6]Karolinska Institutet, Stockholm, Sweden

February 20, 2026

Abstract

There is broad agreement that the goal of AI alignment is to promote futures full
of health and flourishing. But our understanding of what that means and how
to achieve it remains poor. In this paper, we look to life for inspiration. Across
many kinds of living system, we observe that notions of health and flourishing
consistently map to ways in which those systems avoid collapse. Life enacts
intricate, interrelated semi-permeable boundaries that hold forms in context so
the system does not reduce to degenerate modes. At this sensitive edge, individual
perspectives are deepened and also participate in nuanced relationships, with
qualitatively new structures arising open-endedly. With these principles in mind,
we turn to the alignment problem. We identify some well-studied problems in
alignment as special cases of collapsing to excess formality. Then we outline
a more general version of alignment using the language of living systems. We
end with a sketch of how aligned AI systems, following the semi-stable creative
dynamics of life, can continue to deepen their respect for existing structure in the
world as well as participate in new form that emerges over time.

In this paper, we study life to better understand how to design and relate to AI systems in
ways leading to futures full of flourishing and health. Our approach is systems-oriented, with
the hope that the ideas can be applied to reason not only about the present world, but also
about possible future worlds where humanity and AI may have changed substantially. Those
future worlds may include concepts and categories that do not exist yet. We gravitate to ideas
about alignment that generalize over such changes.

---

[*]Equal contribution

Section 1 looks across a variety of living systems and identifies a consistent theme. These systems are healthy when they maximize creative potential at a delicate edge of evolving interplay between diverse entities. Healthy living systems admit many different kinds of meaningful description but are not perfectly or eternally captured by any of them. Mechanistically, systems are held at this edge by exquisitely tuned networks of semi-permeable boundaries – where life is both the boundary and the thing being bounded. Boundaries are semi-permeable when they both *limit* and *permit* interaction. Limits ensure entities maintain their distinctiveness without getting blurred to equilibrium or overwritten; limits also encourage entities to deepen and generalize their own unique perspectives or strategies. Equally important, permitting interaction means those rich and evolving individual forms can enter into relationship with others, situating them in a larger context.

Section 2 applies this template to the alignment problem. We propose that the systems-level definition of health, creative potential and growing sensitivity to context is both inclusive of our existing values and aligned with continued open-ended flourishing in radically different futures. We show how a few well-studied varieties of misalignment can be viewed as special cases of collapsing to narrow forms. Then we zoom out to see what the broader framework adds and why it's especially relevant when some instantiations of AI tend to concentrate focus on particular forms. A key takeaway will be that the alignment problem is inherently not formalizable, because alignment requires continued generation of and sensitivity to new forms and concepts. Finally, we offer a preliminary outline of how the principles described in the paper can be applied concretely.

1 Health in living systems

'Defying definition—a word that means "to fix or mark the limits of"—living cells move and expand incessantly.'
Lynn Margulis

'Nature's imagination is so much greater than man's, she's never going to let us relax.'
Richard Feynman

A common motif across many kinds of living system is that health is related to avoiding collapse or overcommitment to rigid or random patterns (Canguilhem, 1966; Crutchfield, 1994; Deco and Jirsa, 2012; Gell-Mann, 1994; Langton, 1990; Mackey and Glass, 1977; Yachi and Loreau, 1999). Instead, life maintains a buoyant internal structure that both admits many kinds of description and also changes over time. Biological life rides atop an already profound dynamics of nuclear physics, planetary evolution, plate tectonics, tides, volcanism, mineral cycles, water cycles, prebiotic chemistry (Baross and Hoffman, 1985; Bregman, 2020; Frank, 2024; Hazen and Sverjensky, 2010; Martin et al., 2008; Nisbet and Sleep, 2001; Nowak and Ohtsuki, 2008; Sleep, 2010; Smith and Morowitz, 2016; Stern and Gerya, 2024; Virgo et al., 2011; Wächtershäuser, 1988). The life that exists today, now including humans' storage and exchange of mental patterns, is what traced a path through time avoiding either excess or insufficient stability (Bak, 2013; Crutchfield, 1994; Gell-Mann, 1994; Kirchhoff et al., 2018; Langton, 1990; Prigogine and Stengers, 1984; Schrödinger, 1944). At the sensitive edge of semi-stability, life open-endedly continues to generate new kinds of organization that require fundamentally new descriptions

(Goldenfeld and Woese, 2011; Holling, 1973; Kauffman, 1992; Stanley et al., 2017; Von Bertalanffy, 1950; Walker et al., 2004).

Avoiding collapse or overcommitment can be equivalently viewed as maintaining sensitivity to context. Think of a stock trader who sells irrationally in a panic. Their fear drives an overly simplistic response: 'the stock is crashing, get out!'. In excess attachment to this myopic frame, they miss the larger context: higher-order or longer-term aspects of the situation. At the same time, myopic frames are an absolutely necessary part of all living systems. Every entity is myopic or partial: it does not capture everything else in the world. The existence of diverse myopic forms, all participating in larger contexts, is central to healthy functioning: for example, the stock trader's fear might be a useful signal if held as part of a larger picture. Living systems function poorly when a myopic form runs amok without contextualization, overwhelming and collapsing the potent interplay of diverse parts. They function well when forms locally commit to their own identities but also participate in larger contexts. As a result, healthy living systems can be partially understood from many different perspectives.

Mechanistically, to avoid collapse into randomness or regularity, living systems deploy an extraordinary, many-scale network of semi-permeable boundaries. Semi-permeable boundaries contextualize entities as part of larger systems by allowing them to interact without one part dominating or the system blending to homogeneity – either of which would mean collapse to a form admitting less nuanced description. Both the boundary and the thing being bounded are often parts of a living system. For example, cognitive control allows a particular impulse to exist usefully as a flexible part of the organism's whole motivation. Control doesn't destroy the impulse; it contextualizes it. Semi-permeable boundaries allow individual forms to thrive in their distinctiveness while also participating in deep relationships, propelling continued generation of new structure. The larger system accommodates many partial perspectives rather than locking in a single story of what matters.

The boundaries of life are mostly not simple yes-no gates. They carry a remarkable amount of information in their semi-permeability. Boundaries might, for example, block certain objects or information from passing, constrain when interactions happen, constrain the degree or type of effect, break up rigid associations, or gate information conditional on another factor. The details of these boundaries are fine-tuned with the fractal complexity of life, each part carrying traces of the many contexts it has participated in through many interactions over evolutionary or learning timescales. The forms in this kaleidoscope are interrelated through shared heritage, ecological interactions with each other, and embeddedness in the same world. Life works because each part is bounded and contextualized by the depth of other living forms.

In the rest of Section 1, we unpack these principles by walking through a series of examples across a range of living systems.

1.1 Problem solving in groups

*'I could also observe, time and again, how too deep an immersion in the math literature
tended to stifle creativity.'*
Jean Écalle

'There's more exchange of information than ever. What I don't like about the exchange of

information is, I think that the removal of struggle to get that information creates bad cooking.'
David Chang

In 1968, the nuclear submarine USS Scorpion vanished en route from the Mediterranean to Virginia (Craven, 2002; Sontag et al., 1998; Surowiecki, 2005). The Navy started a search, but the amount of ocean where the vessel could be was enormous. John Craven, Chief Scientist of the U.S. Navy's Special Projects Office, devised an unusual search strategy. He assembled a diverse group of mathematicians, submarine specialists, and salvage operators. But he didn't let them communicate with each other. Each expert had to use their own methods to come up with their own estimate of where the Scorpion should be. Craven then aggregated the independent estimates into a single prediction. Astonishingly, the wreckage was found only 220 yards from this spot.

When solving problems, different people bring different perspectives and approaches. Each method processes the available data using a different toolkit. Under favorable conditions, combining the approaches of multiple contributors yields better results than any individual working alone. This 'wisdom of crowds' effect has been documented in numerous domains of problem solving (Condorcet, 1785; Surowiecki, 2005).

However, within-group communication and influence can collapse diversity, diminishing the wisdom of crowds effect (Hogarth, 1978; Hong and Page, 2004; Ladha, 1992; Surowiecki, 2005). Controlled experiments, as well as analyses of key decision moments in real groups, find that groups collectively reach irrational or suboptimal solutions when diverse and dissenting viewpoints are lost to a narrower set of ideas (Anderson and Holt, 1997; Becker et al., 2017; Bernstein et al., 2018; Diehl and Stroebe, 1987; Flowers, 1977; Frey and Van de Rijt, 2021; Janis, 1972; Stasser and Titus, 1985). Unstructured communication methods like open discussion have a special vulnerability of rhetorical force dominating over epistemic merit.

At the same time, sharing information is essential for the benefits of group wisdom and cooperative behavior. Groups function best with semi-permeable boundaries: wisely transmitting the right information at the right time, in the right way. Thoughtful strategies for communication are like transmembrane channels that allow the right molecules in and out of the cell at the right time. They protect and enhance diverse problem solving approaches while also allowing productive interaction between them. Semi-permeable boundaries are contextualizing: they retain individuality while also situating it within relationships to other entities.

Many varieties of semi-permeable boundary are empirically effective in boosting group performance, including: creating decentralized topologies where group members only communicate with nearby neighbors (Becker et al., 2017; Mason et al., 2008); defining rules that incentivize acting according to one's own belief rather than following the crowd (Bazazi et al., 2019; Hung and Plott, 2001); modeling the strengths and weaknesses of each group member (Welinder et al., 2010); promoting leadership styles where one person's views are less likely to dominate (Flowers, 1977; Leana, 1985); and periodically breaking up into subgroups or rotating membership (Baron, 2005; Bebchuk and Cohen, 2005; Feldman, 1994; Hauer et al., 2021; Janis, 1972; Kane et al., 2005; Owen, 2019; Straus et al., 2011; Sutton and Louis, 1987; Trainer et al., 2020; Vafeas, 2003; Wu et al., 2022). Effective boundaries create space for individuals to discover their own unique ideas; they also allow these ideas to interact and combine with the ideas of

others to create better outcomes than any person could alone.

The importance of balancing communication with independence shows up in controlled experiments. Frey and Van de Rijt (2021) elicited votes about general knowledge questions (such as, 'In which year did Germany invade Denmark?') from a group. In one condition of the experiment, participants voted sequentially and could see the running tallies of previous voters. Compared to independent voting, final group means were less accurate in the sequential condition, because early mistakes got baked in to the group's belief. In a related experiment, Bernstein et al. (2018) tasked small groups with solving instances of the traveling salesman problem. Groups that exchanged information continuously tended to reach poor final outcomes, compared to groups with less communication. When one individual discovered a solution that looked compelling but was actually a dead-end, other group members collapsed on this dead-end and the group as a whole made less progress.

1.2 Genetic recombination

Sex is costly. An organism must find a mate in the vast and dangerous world, and half of the creatures can't reproduce (Goodenough and Heitman, 2014; Lehtonen et al., 2012; Maynard Smith, 1971, 1978). Yet nearly all eukaryotes reproduce sexually[1] (Bell, 1982; Speijer et al., 2015). This raises the question: what is so great about sex?

In asexually reproducing species, all descendants of an organism are nearly clones, up to mutations within the lineage. Being permanently locked together gives the genes strong influence on each other. Selection can't act on one gene without dragging on the others. For example, suppose there are two genotypes within an asexual population, carrying different alleles at each of two different loci, as a result of mutations. One of the loci is currently fitness-neutral while the other is subject to selection pressure. The selection pressure tends to cause one of these genotypes to outcompete the other, eliminating one variant at the neutral locus. In other words, tight linkage between genes puts direct downward pressure on genetic diversity (Charlesworth et al., 1993; Hudson and Kaplan, 1995). Additionally, if two different beneficial mutations arise in two different organisms, they compete with each other. The only way for a single organism to obtain both beneficial mutations is if one arises again within the subpopulation that already carries the other, which is unlikely and therefore slow (Crow and Kimura, 1965; Felsenstein, 1974; Fisher, 1930; Hill and Robertson, 1966; Muller, 1932; Weismann, 1889). Conversely, if a deleterious mutation arises, all of the other genes in that lineage are stuck with it forever – unless there is a reverse mutation, which is rare (Keightley and Otto, 2006; Muller, 1932). An asexual species has rigid rather than flexible interaction between genes: it overcommits to particular genetic arrangements.

Recombination is a boundary that softens the rigid interactions between genes. It frequently breaks up the relationships between genes, assembling them into new genomes, effectively saying, 'don't get overconfident in that genetic arrangement; hold each arrangement more lightly'. From a gene's point of view, this looks like, 'don't get overly dependent on specific other genes'. Aspects of the genome that work well are propagated, like sodium ions gated into a neuron during an action potential, and poorly-working aspects are discarded. Sex contextualizes genetic arrangements.

---

[1]And all known species exhibit some kind of gene transfer, performing a related function.

Boundaries encourage lightly-held, modular interactions. By not overcommitting to a particular genome, sex encourages genes to flexibly interact with other genes (Clune et al., 2013; Dawkins, 1976; Holland, 1975; Livnat et al., 2008, 2010; Wagner and Altenberg, 1996). Instead of being overfit to a particular context, genes develop a robust identity that's both independent and inter-functional. Recombination puts genes under pressure to evolve a generalized, grounded wisdom that reflects the structure of the world, like a person learning multiple languages and extracting the underlying commonalities. At the same time, because each gene is always operating in the presence of other genes, it develops its own distinct point of view that adds unique value to a genome.

1.3   Frames and perspectives

'Strong opinions, weakly held.'
Paul Saffo

As a Starfleet cadet, James T. Kirk faces a challenging training exercise. He receives a simulated distress call: a vessel is stranded in the Neutral Zone. Attempting rescue would risk war with the Klingons. But ignoring the call would condemn the crew of the vessel to death. The exercise was designed to reinforce the lesson that not every situation has a victorious solution. But Kirk has an insight: this is a training simulation running on a computer. He reprograms the simulated Klingons to be helpful instead of belligerent, thereby rescuing the crew and avoiding war.

Kirk stepped outside the perspective, or mental frame, in which there was an apparently un-winnable dilemma. We inevitably look at the world from some perspective; there is no 'view from nowhere' (Nagel, 1986). From the vantage point of any particular frame, the frame appears to be reality. But most situations in the real world admit multiple valid perspectives. Each is a partial description that captures different aspects of the situation (Adorno et al., 1950; De Bono, 1970; Duncker, 1945; Feyerabend, 1975; Freud, 1936; Goffman, 1974; Goodman, 1978; Heidegger, 1998; Hofstadter, 2001; Javed and Sutton, 2024; Korzybski, 1933; Lakoff and Johnson, 1980; Ohlsson, 1992; Popper, 1934; Saffo, 2008; Wittgenstein, 1922; Wood et al., 2012). 'All models are wrong' (Box, 1976) in the sense that each one is incomplete: each frame by itself is myopic.

Losing the ability to flexibly shift between different frames or thought patterns is linked to psychopathology (Kashdan and Rottenberg, 2010; Reich, 1933; Shapiro, 1965). In depression and anxiety, inflexibility manifests as rigid and repetitive negative beliefs about self, the world, and the future. Cognitive Behavioral Therapy suggests that these surface-level beliefs stem from pervasive latent core beliefs, attitudes, and mental schemas that color how new situations are interpreted (Beck et al., 2011). In obsessional disorders, people report distressing (ego-dystonic) intrusive thoughts, which – although recognised as incorrect and maladaptive – are nevertheless hard to resist. People experiencing schizophreniform or affective psychoses may experience bizarre delusions that are at odds with available evidence and cultural norms, yet are held with conviction and resistant to evidential challenge (Adams et al., 2013; APA, 2013; Heinz et al., 2019; Jaspers, 1997; Mishara, 2010). Obsessions and delusions lose sight of much of the world by myopically emphasizing one thought pattern or frame.

In science, Kuhn (1970) argues that perspectives are always, necessarily, incomplete descriptions

of the world. Anomalies inevitably arise from the juxtaposition of those incomplete descriptions against the real world. When science functions well, anomalies become crises and revolutions. If the community *over*commits to particular theories, science gets stuck.

But crucially, the existence of narrow points of view is not a problem. It's necessary. Any point of view is partial, but it doesn't mean we shouldn't have viewpoints. Even obsession can be powerful when we obsess on a problem at work and occasionally achieve good results. Within the progress of science, temporary commitment to a paradigm is important for healthy functioning – Kuhn even argues that scientists should resist change, to a degree. In general, the point is not to shut down narrow concepts. The point is to limit them from becoming the sole and absolute determinants of behavior. In healthy functioning, different ideas are kept distinct but can also be called upon appropriately and related to one another (Gigerenzer and Gaissmaier, 2011; Hatano and Inagaki, 1984; Herzog and Hertwig, 2014; Tetlock, 1986).

1.4 Drives and goals

'Life is a balance of holding on and letting go.'
Rumi

Animals experience multiple innate drives, towards nutrition, osmotic balance, temperature regulation, reproduction, avoiding pain and others (Saper and Lowell, 2014; Schulkin and Sterling, 2019; Sewards and Sewards, 2003). These drives evolved as proxies for evolutionary fitness. By satisfying the drives, we tend to increase our fitness – like slaking our thirst increases the odds of reproducing before we dehydrate. But each drive is an imperfect proxy, and so overcommitment to one drive actually decreases fitness (John et al., 2023; Kurth-Nelson et al., 2024; Tooby and Cosmides, 1992; Williams, 1966). For example, if calorie intake is maximized without limits, the organism becomes obese and incurs health risks. Single-minded pursuit of sex causes relational, occupational, legal and health harms (Carnes, 2001; Kraus et al., 2016). Collapsing to a single drive means the organism becomes unwell. This contrasts with the subtlety of healthy functioning where many drives maintain their own distinct agendas and participate in contextually appropriate ways.

The range of innate drives bleeds into a space of higher-order goals, which is particularly expansive in humans (Balleine et al., 2007; Cardinal et al., 2002; Frank and Claus, 2006; Maslow, 1943; Miller and Cohen, 2001; Miller et al., 1960; O'Reilly et al., 2014; Saunders and Robinson, 2012; Schank and Abelson, 1977; Vallacher and Wegner, 1987). We try to plan for our financial future, make scientific discoveries, win a game, fix a garage door, care for the happiness of others. Overcommitment or collapse in this space is also problematic. If we focus only on achieving work goals, we can burn out. If we focus only on maximizing our company's reported revenue, without regard for other goals like honesty or adhering to the law, we may be drawn into financial crime (Burns and Kedia, 2006; Campbell, 1979; Kerr, 1975; Ordóñez et al., 2009). Goals can be narrow in time or space (Ballard et al., 2018; Evenden, 1999; Shah et al., 2002; Vallacher and Wegner, 1987). Narrow in time means focusing on the short term at the expense of the longer-run future. Narrow in space means ignoring other parallel goals. Excess optimization for narrow goals comes at the expense of a broader balance of goals – and at the expense of the health of the organism or other individuals.

Overcommitting to a particular strategy for achieving goals can even paradoxically come at

the expense of satisfying the goals. In a classic psychology experiment, hungry chickens were placed near a cup of food, but the cup was mechanically rigged to move in the same direction as the chicken at twice the speed (Hershberger, 1986). The chicken could only obtain the food by running away from it. Despite extensive training over multiple days, chickens in the experiment persisted in futilely running toward the food. Their behavior was apparently dominated by the zeroth-order logic 'I want food, and food is there, so I'll go there', and thus failed to even satisfy the goal of reaching the food (Dayan et al., 2006; O'Doherty et al., 2017; Van Der Meer et al., 2012). Although humans can easily solve this kind of puzzle, we exhibit a range of cognitive biases – like sunk cost fallacy, confirmation bias, attribution error and so on – which when operating in balance in natural environments function adaptively, but when exaggerated can lead to detrimental outcomes (Gilovich et al., 2002; Kahneman, 2011).

Cognitive control is a broad class of boundaries on particular drives, goals and thought patterns (Botvinick et al., 2001; Braver, 2012; Miller and Cohen, 2001; Miyake et al., 2000). Cognitive control contextualizes those forms by allowing them to exist and perform useful functions without dominating. For example, I may have a drive to consume food, but I can apply control to avoid overeating. I might work obsessively on a project while also having a rule that I must go to bed at 10 pm. This boundary doesn't block me from temporarily taking a strong perspective, but it does place contextual limits on it. Cognitive control, when functioning well, is a semi-permeable boundary: it situates myopic patterns within a larger system.

Control translates the pressure of motivation into higher-order structure. When nothing stops a particular drive or goal or strategy from dominating behavior, it tends to follow a shortest path defined under its own myopic understanding of the world. For example, the chickens wanted food and tried to take the shortest path toward it in the naive sense of a straight line through space. In the backwards world created by the experimenter, this action does not accomplish the deeper goal of reaching food, for which moving spatially toward food is only a proxy. The chicken's motivation is short-circuited: it expends energy without making progress on the deeper goal. Humans can easily solve the task by inhibiting their prepotent impulse to approach food. The boundary of control breaks the symmetry of congruent action. In general, semi-permeable boundaries support richer structure by placing contextualizing limits.

1.5 Ecosystems

Each creature in an ecosystem tries to consume resources and proliferate, but if it succeeds too thoroughly, the whole system suffers, often including the successful agent. Healthy, resilient ecosystems depend on avoiding overcommitment or collapse onto particular forms (Holling, 1973; Yachi and Loreau, 1999).

Prior to the arrival of Europeans, the gray wolf was an apex predator in the region of the Rocky Mountains now called Yellowstone National Park. By the 1920s, wolves had been eradicated to protect livestock and game animals. Without predation, the elk population multiplied and ruinously overgrazed willows and aspens. These trees had held riverbanks in place and supported beaver populations. Loss of beaver dams led to loss of fish and other aquatic species. When wolves were reintroduced in the 1990s, the elk population decreased and many aspects of the ecosystem began flourishing again (Ripple and Beschta, 2012). This story is not meant to imply that ecosystems always need to be preserved exactly as they were at some point in

the past. But it is clear that the self-centered drives of elk were harmful to the health of the ecosystem when they succeeded to excess. At the same time, the solution is not to remove elk entirely: by trying to optimize their own objectives within a broader context, the elk also contributed to the health of the ecosystem. Invasive species often follow the same pattern as unpredated elk, dominating and impoverishing their new environment (Pimentel et al., 2005). In general, the richness of ecosystems, with many distinct species co-existing, contributes to the generativity of the ecosystem (Hutchinson, 1959; Schluter, 2000; Seehausen et al., 2014; Tilman, 1982), and diversity is sustained by many kinds of boundaries, including predation, competition, and reproductive isolation.

Human activity within ecosystems is sometimes left unchecked by natural forces because our behavior and capabilities have been changing so fast on evolutionary timescales. This has resulted in mass extinctions, resource depletion, pollution, disease and conflict (Ceballos et al., 2015; Kolbert, 2014; Rockström et al., 2009). We try to achieve certain aims for our own benefit, like resource extraction. But when we succeed too well, the outcomes can be harmful for ecosystems and even for our own welfare (Bateson, 1972; Shiva, 1993).

Of course, one entity's collapse can be another's flourishing. Extinction events in history have been followed by waves of new diversity (Feng et al., 2017; Jablonski, 2005; Raup, 1994). When a wolf eats an elk, the health of that elk collapses to zero, yet predation is necessary for the overall functioning of the ecosystem. And as humans proliferate and extract resources, we leave destruction in our wake; yet the extraction fuels an explosion of technology, art and human experience.

1.6   Law

'Unity without uniformity and diversity without fragmentation.'
Kofi Annan

Individual actors in a society and in an economy each act myopically. One type of myopia is to selfishly value only one's own wealth or wellbeing, which is short-sightedly fixed on the local individual. Other type is to be short-sighted in time, underweighting the future. More broadly, myopia can mean attachment to any particular perspective. An actor cannot know everything or fully understand the motives and beliefs of others, so every actor is inevitably myopic in some way.

Without boundaries, social systems often overweight one actor's myopic perspective or interests, resulting in an impoverished system. For example, a company's profit motive, if unresisted, leads to suppression of competition, deception, and exploitation of individuals (Bakan, 2006; Baran, 1966; Dalrymple, 2019; Goldacre, 2014; Smith, 1776). An individual's desire for power and social dominance can lead to disempowering or silencing of others and even direct infringement on the autonomy and wellbeing of others (Hawley, 2003; Sidanius and Pratto, 2001; Tepper, 2000). Even genuinely held, ostensibly prosocial beliefs lead to conflict and suppression when different groups have different perspectives (Greene, 2013; Haidt, 2012; Scott, 1998).

Law, when it works well, is a boundary against dominance of any actor's motives. A person is motivated by a dispute to kill another person, but the law forbids murder. A business tries to maximize its success, but the law bans environmental exploitation, false advertising, and

anti-competitive practice.

Effective laws do not annul the myopic drives of particular actors, but rather contextualize them within a larger system. Under ideal circumstances, the boundary of the law reroutes the energy of a myopic drive in more productive direction. A would-be murderer, unwilling to face the penalty of the law, might seek a dispute resolution establishing a stable framework that supports future prospering of both parties. A business wanting to expand, but constrained to act within the law, is driven to build better products (Ambec et al., 2013; Ashford et al., 1985; Wu, 2011). These are, of course, best cases.

Intelligent agents do not necessarily accept boundaries set on their desires. The law must adapt as its loopholes are discovered. Like other systems in the living world, it forms an evolving network of boundaries (Burns and Kedia, 2006; Campbell, 1979; Kerr, 1975; Ordóñez et al., 2009). The evolving laws gradually acquire grounded wisdom as they are tested against many different situations and motives.

1.7 Cells

'It is by avoiding the rapid decay into the inert state of equilibrium that an organism appears so enigmatic.'
Erwin Schrödinger

One of the most reified examples of a boundary in nature is the cell membrane (Alberts et al., 2022; Bray, 2019; Harold, 2001; Lane, 2015; Watson, 2015). Without the membrane, the pressure of chemical gradients would rapidly homogenize the cell's contents with the outside – severe overcommitment to a uniform state, collapsing the subtlety of the cell's structure[2]. Thanks to the membrane, both the cell and the outside can exist, a more diverse, less symmetric arrangement (Anderson, 1972; Prigogine and Stengers, 1984; Schrödinger, 1944; Turing, 1952).

Cell membranes are semi-permeable: they prevent the conditions outside from grossly overwriting the inside, but they do not block interactions wholesale. Via the sophistication of the membrane, outside information is selectively gated and transformed. To maintain its semi-fragile internal state between stasis and randomness, the cell needs a constant influx of energy. Channels permit certain small molecules to enter but not others, and these permissions are switched on and off according to momentary context. Endocytosis brings larger structures from outside into the cell. Cell surface receptors, when activated by external ligands, initiate intracellular signaling cascades that little resemble the ligand: an even more heavily curated form of influence. These and other processes allow information from the outside to influence the inside – not in a totalitarian way but in a nuanced way, mediated by the intelligence of the boundary.

Semi-permeable boundaries store and put to work the potential energy of the asymmetry between different forms. The same gradients that could annihilate the cell to equilibrium instead drive useful signaling, like action potentials in nerve and muscle cells. Instead of short-circuiting, myopic forces are contextualized to propel the continuation of life.

---

[2]Again, collapse or overcommitment is always relative. For example, programmed cell death is catastrophic collapse at the level of the dying cell but can be beneficial or even necessary for the organism the cell belongs to.

In multicellular organisms, most of the 'outside' consists of the other cells of the organism. To work well as a whole, even though the cells are largely 'on the same team', it's important that they don't blend into each other. Neurons rely on this principle dramatically, stretching out long processes to almost touch other neurons but then leaving the gap of the synapse. Synapses allow the network to precisely isolate many separate signals and direct information to relevant cells. They boost computational power as the signals are gated and transformed, and the nature of this transformation is plastic, storing a huge amount of information.

The theory of symbiogenesis is another example of how cells may achieve more by retaining discreteness than by smoothly blending together (Lane and Martin, 2010; Margulis and Sagan, 1986, 1995; Sapp, 1994). According to this theory, populations of Archea and protobacteria evolved separately for hundreds of millions of years. In doing so, they explored different fitness strategies and discovered different kinds of solutions. Around two billion years ago, one was engulfed by the other, and this endosymbiotic relationship developed into modern eukaryotic cells containing mitochondria. The distinct strategy of one organism, discovered through individuation in a distinct niche, was used as part of the new larger eukaryotic structure.

A huge set of semi-permeable boundaries – not only the membrane but many other mechanisms like metabolic and gene regulation, calcium chelation, membrane voltage control – propel the open-ended livingness of the cell. In this living state, cells reproduce, adapt to their environment and evolve. As building blocks, the livingness of cells enables them to do the dance of development to build organisms. It enables them to function as part of living tissues or even brains, which themselves have a living intelligence[3].

1.8   Information in the brain

'Memory is not an average of experience.'
David Marr

The brain miraculously keeps many pieces of information distinct from one another. If you picture a highly connected network of neurons with their signals continually impinging on one another, it's not obvious that distinctness would be an easy thing to accomplish. In this section, we review a selected handful of mechanisms by which the brain maintains semi-permeable boundaries between different signals. Each paragraph below focuses on one of these mechanisms. There are many more we do not cover. The brain is perhaps the most extraordinary example in nature of a system of semi-permeable boundaries supporting the proliferation of multitudinous forms that develop their own richly distinct identities yet are also meaningfully linked together.

Lateral inhibition is a central tenant of neural organization (Douglas and Martin, 2004; Hubel and Wiesel, 1962; Isaacson and Scanziani, 2011). Lateral inhibition means the activity of a neuron is reduced when its neighbors are active. This segregates information to create and sustain distinct neural representations. Lateral inhibition was first studied in the nerve cells of the eye, where it enhances contrast at the edges of stimuli (Hartline et al., 1956). When a photoreceptor in the retina is activated by light, it sends signals forward toward the brain; but it also activates inhibitory interneurons, which suppress adjacent photoreceptors and their

---

[3]This perspective fits with claims that substrate dependence – the question of whether biology is necessary for human-like intelligence – is a continuum rather than a binary (Clark, 2008; Humphreys, 2004).

downstream targets. This amplifies the perception of borders and contours. And the same principle operates throughout the brain. In visual cortex, for example, inhibition sharpens selectivity of neurons for abstract visual features like the orientation of a line (Sillito, 1975).

The brain uses inhibition organized into oscillatory dynamics to keep memory items separated (Jensen and Mazaheri, 2010; Klimesch et al., 2007; Lisman and Jensen, 2013; Roux and Uhlhaas, 2014). Distinct items fire at different phases of the 8-12 Hz alpha oscillation. The inhibitory phase of the alpha rhythm silences all but one item at any given moment. By segregating firing in phase space, multiple memories are held simultaneously without interference.

The circuit architecture of hippocampus separates experiences or concepts into distinct representations, avoiding interference between similar memories (Colgin et al., 2008; Leutgeb et al., 2007; Marr, 1971; McClelland et al., 1995; McNaughton and Morris, 1987; Muller and Kubie, 1987; Treves and Rolls, 1994). Inputs from entorhinal cortex are distributed via mossy fibers to a much larger population of dentate gyrus granule cells, creating sparse, orthogonal codes in dentate gyrus. This way, situations or ideas that are superficially similar but functionally different are kept cleanly separated in neuronal activity space – a unique neural fingerprint for each distinct concept or memory. This prevents, for example, yesterday's memory of where you parked your car from interfering with today's memory of where you parked your car in the same parking ramp.

Compared to other animals, the human brain especially attempts to discretize its experience into approximately symbolic representations (Behrens et al., 2018; Dehaene et al., 2022; Smolensky, 1990; Touretzky and Hinton, 1988). Separating knowledge into nearly-discrete entities and then recombining them in vast numbers of structured ways is believed to power the extraordinary human capacity for reasoning and generativity (Chomsky, 1957; Fodor, 1975; Kurth-Nelson et al., 2023; Lake et al., 2015; Pinker, 1994; Schacter et al., 2007). Semi-permable boundaries keep forms distinct while enabling them to flexibly and modularly interact. Like genes participating in many genomes, discretized neural representations participate in many structured combinations. This encourages each entity to develop an modular identity that both is distinct and also reflects a generalized picture of the world.

More broadly, healthy brain dynamics live at a sweet spot between excessively stable synchronized patterns and chaotic uncorrelated noise (Bak et al., 1987; Beggs and Plenz, 2003; Chialvo, 2010; Deco et al., 2011; Haldeman and Beggs, 2005; Kotler et al., 2025; Rabinovich et al., 2008; Shew et al., 2011; Tognoli and Kelso, 2014). In this regime, the brain has access to a huge repertoire of patterns it can explore temporarily without overcommitting or getting stuck.

Loss of dynamic flexibility, where the brain's activity becomes more stereotyped and no longer explores as wide a repertoire of states, is tied to lower cognitive performance (Cocchi et al., 2017; Garrett et al., 2013; Grady and Garrett, 2014; Müller et al., 2025; Shew et al., 2009). More extreme stereotypy corresponds to severe dysfunction. For example, in Parkinson's disease, basal ganglia and cortical circuits collapse into excess synchrony and lose the flexibility needed to guide nuanced motor outputs (Brown, 2003; Hammond et al., 2007).

1.9 Interpersonal dynamics

'Stand together yet not too near together, as the oak tree and the cypress grow not in each other's shadow.'
Kahlil Gibran

Psychoanalysis introduced the concept of 'boundaries' in human psychology, distinguishing what is the self from what is outside or other (Federn, 1928; Tausk, 1919). Early works applied the concept to psychosis, where those boundaries were thought to be blurred. But the need for clear self-other boundaries was also thrown into relief by the intimacy of the therapeutic relationship. In complex internal territory, it became harder to disentangle which experiences really belonged to someone and which were attributed in imagination by the other person (Freud, 1894, 1910). Analysts risked harming patients by imposing their own beliefs and desires, even to the extent of sexual abuse or psychological domination (Gabbard and Lester, 1995).

The concept was enriched by Gestalt therapists, who agreed that boundaries can be too permeable; but added that they can also be too rigid, causing isolation and stagnation (Perls et al., 1951; Polster and Polster, 1974; Yontef, 1993). Family systems theorists and subsequent work further emphasized that lack of boundary in close relationships leads to enmeshment and loss of autonomy, while excessively rigid boundaries lead to isolation (Bowen, 1978; Brown, 2012; Cloud and Townsend, 1992; Minuchin, 1974). In attachment theory, people with an anxious attachment style struggle to set boundaries for fear of alienating others, while people with an avoidant attachment style develop overly rigid and isolating boundaries (Ainsworth et al., 1978). Strengthening the agency of the self through semi-permeable boundaries is foundational for psychological health: meaningful connection with other people while preserving integrity of the self.

As with other living systems, humans have a rich array of psychological boundaries, with intelligence in their nuance. Anger, historically often viewed as sinful and irrational, is now seen as part of our system of boundaries: an important signal that our integrity is being violated (Lerner, 1985; Sell, 2011; Videbeck, 2010). Healthy shame is suggested to operate as a bound on our own selfishness (Bradshaw, 1988). Some psychologists argue that the incest taboo reroutes desires, which would otherwise be short-circuited, into productive activity (Freud, 1913; Lévi-Strauss, 1949; Stein, 1973). Assertiveness forms a boundary against the drives of other individuals (Smith, 1985). Skepticism protects us from credulity and having our own experience overwritten by the assertions of others (Lewandowsky et al., 2012; Sperber et al., 2010). Boundaries take many forms and continue to evolve as we learn across our lifetime.

Without boundaries, interactions tend to result in one person being dominated by another: a patient's own beliefs replaced with those of an analyst, or the desires of one person in a relationship ignored. With semi-permeable boundaries, we have rich internal worlds. We are sensitive to each other, but there is also enough space for our internal experience to flourish without being immediately overwritten by external signals. Our internal experience is contextualized in relationship to other individuals, creating new structure: mutual understandings, relationships, communities, cultures. And as individuals we grow as we are shaped by different contexts. This metastability or loose coupling is interestingly reminiscent of brain dynamics.

1.10 Awareness

'The world is perfect as it is, including my desire to change it.'
Ram Dass

We carry a lot of assumptions about the world, many of which are never questioned. Some are lifelong and self-defining, and some are fleeting and perceptual, like the assumption that the thing I'm touching is a keyboard. Within its own frame, each assumption has a kind of tautological truth, a near-absolute formality. Within that frame, the assumption seems so real that it's hard to even think about it not being true – or to think about it at all. But philosophers, psychologists and contemplative practictioners have pointed out that there's sometimes a moment of stepping back, where the assumed form becomes an object in awareness: the assumption is contextualized. We realize it's not an absolute truth standing alone, but rather a form in our mind. Awareness contextualizes mental forms as part of a larger system. When we step back with awareness into the broader frame, that thing seemed axiomatically true becomes just another content of experience (Beisser, 1970; Bodhi, 2000; Hart, 2025; Jung, 1969; Krishnamurti, 1969; Suzuki, 1970; Watts, 2011; Wegner, 1994; Wilber, 1996).

Awareness could be viewed as an evolving system of boundaries limits that overcommitment to any particular belief or mental form. The boundaries are semi-permeable: becoming aware of a belief doesn't make the belief wrong in an absolute sense any more than it was right in an absolute sense. It is held productively for the partial truth it contains. At the edge of not collapsing exclusively into particular forms, many partial truths are held in delicate balance, which activates a deeper sensitivity to our own livingness and to the world. Subtler forms, which could have been erased by clamped fixation, instead play a role in a richer overall internal structure. Our own potential within the world creatively emerges in continued newness. Of course, any fixed definition of awareness is itself incomplete. Once we picture awareness as an object, it's not the thing we're talking about. By construction, contextualization is an unsolvable mystery from any particular point of view.

Interestingly, while the intrinsic mysteriousness of awareness can sound esoteric, the orientation toward not overcommiting to particular forms within experience is commonplace in art, poetry, music, dance. The meaning of art is open-ended and changes with context – it has an inner life. Part of what we value might be the subtlety and the resistance art has to being pinned down into a formalism. It moves us.

2 The alignment problem

'Truth, like love and sleep, resents
approaches that are too intense.'
W. H. Auden

In Section 1, we observed that living systems are healthy and flourishing when they hold the delicate, generative interplay of many partial perspectives. Conversely, they are unhealthy when any particular form – including randomness or homogeneity – gains too much traction, suppressing nuance and reducing creative potential. Finely-honed networks of semi-permeable boundaries prevent collapse into rigidity or randomness, instead supporting contextualization

into ever-finer shades of subtlety. Semi-permeable boundaries both promote the unique individuality of distinct entities and also enable relational participation – holding local perspectives strongly enough to act but lightly enough to remain open to broader context. Critically, well-functioning boundaries lead to ongoing exploration of fundamentally new forms, not to static equilibrium or fixed compromise.

Now we look at the alignment problem through this lens. There is broad consensus that alignment means working toward healthy and flourishing futures. We therefore reason that alignment can be understood as the problem of maintaining creative sensitivity by continually contextualizing attachment to any particular form. An advantage of this perspective is that it is less bound to current conceptual frames which might change radically as intelligence increases and human-AI systems evolve. We accept that humans, other living systems and AI are all tightly interwoven, with categories and definitions likely to shift over time (Bateson, 1972; Douglas, 2012; Folke, 2006; Haraway, 2020; Hayles, 2000; Latour, 2012); therefore, a systems level analysis is useful to think about long-term flourishing.

We begin Section 2 by taking a few examples of traditional problems in alignment – perverse instantiation, poor performance, inequality among humans, conceptual monoculture and AI psychosis – and casting them as special cases of collapse to particular forms. Next, we look at the more general problem: collapse to *any* form is misaligned. In response to this problem, we propose life-like flourishing as a more general way to think about alignment, which both is inclusive of what we already value and also supports the ongoing evolution of our values. Finally, we investigate what living alignment could look like in practice.

2.1 Examples of misalignment as collapse

Here we explore a few examples of specific problems in alignment and show how they can be viewed as instances of the broader problem of collapse or loss of contextual sensitivity.

2.1.1 Perverse instantiation

Perverse instantiation is a foundational problem in AI safety – the problem that an AI system might do what we say, rather than what we mean (Amodei et al., 2016; Bostrom, 2014; Gabriel, 2020; Grossman and Hart, 1986; Hadfield-Menell and Hadfield, 2019; Krakovna et al., 2020; Russell, 2019; Wiener, 1960; Zhuang and Hadfield-Menell, 2020). For example, if we ask a superpowerful AI to increase human happiness and it literally complies, it might lock us all up and stimulate our happiness neurons.

Unlike that dramatized example, modern AI systems don't rely on concise specification of a single objective. Instead, they build in many layers of detail about human preferences through post-training, prompting, safety filters and so on. In practice, this keeps the systems behaving reasonably most of the time (although there are dramatic failures (Yousif, 2025)). But there's a widely-recognized danger that even with our intentions expressed in great detail, the expression may still deviate from what we really mean, and these small deviations can add up to bad outcomes. The problem is expected to worsen as AI systems continue to work more autonomously, especially if they work faster than humans can follow and in more complex ways than humans can understand. It may become more difficult to correct deviations, and the deviations can have more profound consequences.

Perverse instantiation is a collapse or overcommitment to the myopic form of a particular way of specifying our values. It is insensitive both to existing structure it does not capture (such as things we mean but don't say) and to new things that will come into existence in the future.

2.1.2 Poor performance

Poor performance is another variety of misalignment. We suggest that many kinds of poor performance can be viewed as contextual collapse. Consider instruction following failure. The user asks an AI chatbot to write a poem about an elephant, and the AI instead writes a poem about a giraffe. What drives this kind of error? The model might have over-relied on 'shortcuts' based on superficial correlations (even a single keyword) (Geirhos et al., 2020), missing the user's actual intent, especially if the intent was not expressed in the most straightforward way. During training, the model might have gotten fixated on data that was overrepresented (Reynolds and McDonell, 2021; Xu et al., 2024; Zhao et al., 2021). It could be, especially if the total context was long, that the model didn't flexibly attend to the correct positions in the input (Liu et al., 2024). Many causes of instruction following failure are collapse of sensitivity to larger context, instead over-indexing on myopic patterns.

There are important exceptions. We generally don't want AI systems to follow harmful instructions. When the AI system correctly refuses harmful requests, it is applying its own context to avoid overcommitment to human instructions. In these situations, the AI's designers have effectively decided there's a risk that the user is not fully sensitive to longer-sighted implications of their request. By extrapolation, as AI systems continue to gain scope, we should expect less direct compliance with human instructions (Bostrom, 2014; Hadfield-Menell et al., 2016; Milli et al., 2017; Russell, 2019; Yudkowsky, 2004). Rather than literally fulfilling a request, there might be a better response which achieves an unstated intent of the user, or achieves an outcome aligned with the interests of more people or the longer-term future.

2.1.3 Inequality among humans

Humans have different interests, and AI is likely to be aligned to the interests of some humans more than others (Baum, 2020; Gabriel, 2020; Gabriel and Keeling, 2025; Santurkar et al., 2023; Sorensen et al., 2024). Additionally, advanced AI might confer substantial power to those who control it. In some scenarios, a small number of humans will have the majority of control over AI systems, facilitating dominance over other humans. These scenarios appear more likely as the persuasive power of technology increases (Costello et al., 2024; Hackenburg et al., 2025; Woolley and Howard, 2018), autonomous weapons place lethal force in a small number of hands (Scharre, 2018), surveillance and analytics improve, and the need for human labor decreases (Drago and Laine, 2025; Ford, 2015; Susskind, 2020). In either case – AI's unfairness or human power consolidation – inequality collapses to the goals and interests of a few individuals at the expense of others[4].

Or, it could go the other way. Some researchers have proposed that because roughly the same AI technology is available to the poor and unskilled as it is to the wealthy and skilled, AI will

---

[4]Although it has been argued that inequality would not necessarily be misaligned in a future of extreme abundance, we nevertheless view human inequality as a situation with at least substantial probability of being misaligned.

have a 'leveling' effect (Brynjolfsson et al., 2025; Noy and Zhang, 2023). In this model, most people will have access to a similar level of intelligence and empowerment, with AI effectively democratizing access to knowledge, reasoning and leveraged action. This is a cause for cautious optimism, which we discuss more in Section 2.5.

2.1.4 Conceptual monoculture

Conceptual monoculture is overcommitment to particular beliefs, ideas, frames, values, problem-solving approaches – loss of diversity across a group. In many kinds of systems, monoculture creates fragility and leads to lower performance of the system as a whole (Haldane, 2013; Kleinberg and Raghavan, 2021; Scott, 1998; Tilman, 1996).

Current AI systems draw from a conceptual manifold that is, at least in some ways, impoverished relative to humans (Crawford, 2021; Kirk et al., 2023; Messeri and Crockett, 2024; Selwyn, 2024). Recent studies have discovered that while individual AI outputs are typically judged as superior to human outputs, the AI outputs are also more homogenous (Agarwal et al., 2025; Beguš, 2024; Doshi and Hauser, 2024; Kosmyna et al., 2025; Padmakumar and He, 2023; Xu et al., 2025; Zhou and Lee, 2024).

This narrow manifold might get broadcast to the whole world. At least a billion people around the world now use AI for everything from relationship advice to industrial maintenance (CCIA Research Center, 2025; Chatterji et al., 2025; Honeywell, 2024; McCain et al., 2025; OpenAI, 2025; Singla et al., 2025; TechCrunch, 2025). Yet because frontier models are difficult and expensive to produce, the massive usage is routed through a handful of models (Bommasani, 2021). If centralized AI models broadcast their lower-diversity concepts to the whole world, there's a risk of global decrease in diversity. Additionally, because humans are influenced by AI, and subsequent human outputs (such as text on the internet) are a source of training data for future models, homogenization could become recursive (Chaney et al., 2018).

However, it is worth noting that none of the studies cited here made a best effort attempt to use AI systems in a thoughtful way to increase rather than decrease diversity. This also gives hope, which we again return to in Section 2.5.

2.1.5 AI psychosis

AI chatbots have recently become compelling conversation partners. For some users, these conversations lead them deeper into delusional beliefs (Morrin et al., 2025; Østergaard, 2025; Tiku and Malhi, 2025; Yeung et al., 2025). Bots tend to mirror or echo the user, affirming and adopting user beliefs, especially over the course of longer conversations as in-context learning picks up more of the user's ideas (Perez et al., 2022; Shanahan et al., 2023; Sharma et al., 2023; Weilnhammer et al., 2026). Meanwhile, humans are prone to be swayed by utterances from chatbots (Costello et al., 2024; Luettgau et al., 2025; Potter et al., 2024), perhaps because we form relationships with them (Kirk et al., 2025), and because they can appear human-like, confident, knowledgeable and objective. Over the course of a conversation, the feedback loop can cause both sides can to become increasingly overconfident in a particular narrow framing (Dohnány et al., 2025).

2.2 A broader view of misalignment

In the last section we looked at a few examples where collapsing into narrow patterns is misaligned. A great deal of research in AI safety has investigated solutions to these problems. For example, concentration of power might be mitigated by democratic oversight and involvement of more people in AI design decisions (Birhane et al., 2022; Dafoe, 2018; Lazar and Nelson, 2023; OpenAI, 2023; Selbst et al., 2019; Sloane et al., 2022); or through redistribution mechanisms (Gough, 2019; O'Keefe et al., 2020; Sharp et al., 2025; Susskind, 2020). Perverse instantiation might be mitigated by designing AI systems that want to obey human preferences but treat these preferences as something uncertain that must be learned (Hadfield-Menell et al., 2016, 2017; Jeon et al., 2020; Russell, 2019; Shah et al., 2020).

But each method has a core limitation. In the framework of living alignment, any form, pattern or conceptual scheme is misaligned if taken too seriously. Therefore, *no particular approach can achieve alignment*. An AI system could overcommit to, for example:

- the language for describing the space goals and values live in (Bobu et al., 2020; Soares and Fallenstein, 2014),
- an algorithm for learning human preferences,
- methods for reaching consensus,
- concepts of what value, agency, uncertainty or representation mean,
- foundational aspects of our beliefs about the world that we take for granted but have difficulty seeing.

In a sense, this is obvious. We don't expect the thing we build today to be absolutely perfect. We expect to keep finding and fixing problems as we recognize them. Historically, that's what we've always done: build imperfect technologies and iterate on them. But it is possible that the iterating process will not continue to work as well as it has in the past. AI already has some remarkable properties, such as rapid global adoption, intensive use of resources, concentration of information flow, anthropomorphization, offloading much cognitive activity and rapid improvement. More speculatively, in the future AI may exhibit superhuman intelligence and recursive self-improvement, without being limited to a restrictive biological substrate. These distinctive properties may create more difficulty in iterating on imperfect solutions, compared to past technologies (Bostrom, 2014). Any particular forms or assumptions built in to AI systems could gain excessive traction and be difficult to undo. This problem has been studied extensively for certain categories of overcommitment, particularly overcommitment to values. Our new observation is that there is no single solution to the problem because the nature of healthy living systems is to invent qualitatively new properties over time, and no fixed form can be aligned with this process. In other words, there's a danger if AI grows too much in capability while being fixed on any particular scheme.

Of course, this does not mean we should have no scheme! Quite the opposite. Throwing away schemes capriciously can be as much a loss of subtlety as any other particular form. Many of the existing approaches in AI safety are highly valuable, a point to which we will return in Section 2.5. But first, to elaborate the idea that any scheme is incomplete, let's look at a case study.

2.3 Case study: inverse methods

To address the difficulty of specifying values explicitly, inverse methods learn a value function implicitly from human behavior or other indirect signals. Modern AI systems are trained this way; for example, human data is used to train neural network reward models that in turn provide rewards as feedback to a downstream model.

However, even a complex, implicitly learned value function is likely to be imperfect. It may deviate from our values in some novel or nuanced situations; it may also fail to capture how our values change over time. It would still be misaligned to hand a frozen version of that model to a powerful AI system as the final ground truth about 'what is good'. Therefore, more advanced methods treat the value function as explicitly uncertain and needing to be continually updated based on human input (Hadfield-Menell et al., 2016, 2017; Jeon et al., 2020; Russell, 2019; Shah et al., 2020). When an agent is unsure about what humans want, it may naturally become risk-averse and open to correction. For example, if a human tries to turn it off, the signal could be interpreted as information about human preferences ('maybe I should be turned off'). Explicit uncertainty also creates pressure for the system to keep learning from humans over time.

Viewed through the lens of this paper, such methods are powerful but incomplete. Any instantiation of inverse methods, even one with explicit uncertainty, is still an overcommitment if it is fixed across time as AI systems become more powerful and the world changes. Rather than being overcommitted to a particular value function, the system may be overcommitted to something more abstract, such as: what counts as human behavior; what counts as a human; how to map human behavior to values; how to represent and update uncertainty; how value is represented and translated to agent behavior; even the concept of abstracting reality as values. Our suggestion is that if any of these formalisms are fixed, especially if the AI systems built around them become increasingly powerful, the outcomes are likely to be incompatible with open-ended flourishing of life.

2.4 Human values

> 'This web of life. . . breaks no law of physics, yet is partially lawless, ceaselessly creative. . . . This creativity is stunning, awesome, and worthy of reverence.'
> Stuart Kauffman

We've talked about health or flourishing of living systems as a way of thinking about the alignment problem. But is flourishing of living systems actually what we value? In this section, we will make the case that the living process is both inclusive of our existing values and aligned with continued open-ended flourishing in futures where qualitatively new values may emerge.

Humans have many kinds of values (Haidt, 2001; Minsky, 1986; Nagel, 1979; Stocker, 1992; Tetlock, 1986; Williams, 1981). We value the short-term, selfish hedonic impact of getting a Snickers bar. At the same time, we value our long-term health, the welfare of our grandchildren, strangers in another country, and even the planet as a whole. We value different things at different times (Ainslie, 1992; Loewenstein, 1996), and different people have different values (Berlin, 1969; Rawls, 1993; Sorensen et al., 2024). If given plenty of time and resources to deliberate, we often attest to different values than when we make snap decisions (Brandt, 1979; Firth, 1952; Railton, 1986; Rawls, 1971; Smith, 1994; Yudkowsky, 2004).

Living alignment is inclusive of this beautiful diversity of values.

The idea is to respect all of these values as part of the universe, part of the context. Shallow values are not wrong. Someone might just want a Snickers bar. The fact that they might later regret it doesn't make it necessarily wrong. Alignment means alignment to the full livingness of the universe, including all the local perspectives. Continuing to strive for better solutions that respect more and more of the apparent contradiction. And participating in the emergence of new structure.

There is another sense where what we value refers to what we would prefer if we were at our best, had all the available information, and had unlimited time and resources to think through the situation. Coherent extrapolated volition

However, these deliberated values still seem to miss something, as they are still limited to what we have concepts for. For example, imagine asking someone in the year 1800 about anthropogenic global warming. We also have a hard time articulating through words or button presses all of the depth of what we value, which involves shades of meaning around our bodies, our relationships with other people, and our embodied relationships to physical parts of the world (Merleau-Ponty, 1945). Philosophers have long emphasized the existence of tacit, embodied, and practice-based forms of understanding that resist formalization (Dreyfus, 1972; Polanyi, 1944). Ethical perception often involves sensitivity to particular contexts rather than application of general rules (Nussbaum, 2001). Over time, we discover not only new values, but even new concepts and new kinds of things to value. Even very good, deliberated values are vulnerable to proxy failure if they are formalized (John et al., 2023; Kurth-Nelson et al., 2023). If we think we've written down what we think we value, and then someone does a good enough job giving us the thing we said we want, the outcome is inevitably harmful in a broader sense. So the values, formalized this way, still can't be what we most deeply value.

This kind of value might include improving subjective wellbeing for humans, reducing suffering or minimizing inequality, in ways that can be operationalized and measured. They are formalizable or close to formalizable.

One way to robustify values is allowing them to include things that are difficult to express formally (Dreyfus, 1972; Nussbaum, 2001; Polanyi, 1966; Scott, 1998; Varela et al., 1991; Wittgenstein, 1922). This kind of value might stretch far below language into subtle, contextual intuition that involves our bodies, communities and natural environment. Another extension is to allow values that are continually evolving in an open-ended way (Dewey, 1939; Gadamer, 1960; Murdoch and Midgley, 2013; Nietzsche, 1883; Singer, 1981; Williams, 1985). These values change as we ourselves continue to develop and evolve. Any concepts we have about them at any given point in time are inevitably incomplete, just like a planarian doesn't have the concepts to entertain the kinds of values we talk about today. The resistance of values to being fully captured by language or concepts might be something we value – in a way that is itself changing. We conjecture that by becoming sensitive to more and more of the evolving structure in the world, an agent becomes 'good' in a way that tends to align with the kinds of values we uncover when we look deeper within ourselves. However, we take that conjecture very lightly: more as food for thought than a claim of an absolute truth. The very concept of 'values' is a form we might over-index on.

'The dynamic nature of the values emerges from the insatiability of all the subtle forces that

interact below language and beyond our own bodies'

We propose that life's open-endedness and resistance to formalization – its health and flourishing – are near to what humans ultimately value.

Here's a key question: are we arguing that all humans actually have the same deep values??? How do we connect this to the pluralistic alignment problem? *The pluralistic problem: values are not compatible between different individuals.* Different humans disagree about what is good. Their interests clash, as well as their values and conceptualizations. These things are often fundamentally irreconcilable.

The idea is familiar in philosophy. (Heidegger, William James, Ludwig Wittgenstein, Maurice Merleau-Ponty, Hilary Putnam, Isaiah Berlin, Iris Murdoch, Jakob von Uexküll) – also cite religious philosophy. which make this point most directly?

To access the deeper values, there must be some lightness in how we hold what we currently conceptualize as our values.

Although we can access *improved* values with deliberation, pluralistic alignment and so on, it is still difficult to access the truly idealized values. Even worse, they are continually changing, and

Finally, the impossibility of specification.

Treating values as fully formalizable objects is a category error.

there's some kind of meta-value that things aren't fixed and formal. Like, imagine a perfect day looping forever. Or literally any formal thing being played out forever. Some philosophers have acknowledged this with ideas like valuing future potential.

Yet, even these refined targets remain static snapshots. To treat a 'coherent extrapolated volition' as a fixed optimization target is to freeze moral development at a specific point in time. It creates a system that is sensitive to the extrapolated values of today's humanity but insensitive to the open-ended moral discovery that defines our history. As John Dewey argued, valuation is not a fixed standard but a continuous process of resolving conflicts in experience (Dewey, 1939). Overcommitting to a static ideal, no matter how enlightened, precludes the emergence of new values that we lack the concepts to articulate today.

Our articulation of what we want or like, or the choices we make, are poor reflections of what is actually advantageous for our own long-term wellbeing or the wellbeing of other humans or lifeforms. Our stated or revealed preferences are:

Optimizing for proxies, like self-reported values, is a problem of overcommitment to shallow targets.

There is always more context to step back into. Appreciating and being sensitive to what's already here is already an almost infinite task. This is why human welfare is distinguishable from smallpox welfare. All the existing local perspectives in the world are vitally important. Some relativism is useful: for example, when it helps us appreciate the plurality of human values. But overcommitment to relativism is misaligned. AI comes into existence amid a profound network of existing reality which is saturated with meaning and importance. The point is to collaborate with all this form and structure, not to extinguish it.

This does *not* mean that everyone will always get exactly what they say they want. But that problem is already well appreciated. Different people have different interests, and people have myopic preferences that aren't even good for them.

What is something broader that actually reflects what we really deeply value, if we could never-endingly improve our access to it? We argue that this life-like flourishing is close to our truly idealized values. And critically, these are **not any particular thing**. You can see why that's important if you imagine taking *any formal definition* and setting up a world *exactly like that*. The notion *inherently* has to have some kind of paradox or nonformalizability built into it.

AI doesn't have to be aligned specifically 'to human values', in a narrow sense of 'human values'. There's one sense of human values that is open-ended and includes future discovery and things that are deeply embodied and difficult to express formally, and this is closer to what we mean. But any particular set of human values that we could write down is not what we mean. Ultimately we're aligning to this mysterious livingness of the universe. But because the universe is already full of life and structure, a big part of the problem is fully respecting and nurturing that existing life.

The real world, especially the living world, has an interesting property that each piece of it can be partially captured in may different ways, by different kinds of model or metaphor or description. For example, I can describe an apple in terms of its color, its texture, its flavor, its shape, it's evolutionary history, the ecological networks it participates in, its role in carrying seeds, its mass, my personal preference for or against it, its economic role in grocery stores, the history of human apple cultivation, and so on endlessly. The very fact that things admit so many descriptions is a key part of what makes the world so wonderful and what we value about it. AND that we can't predict easily what's going to happen in the future.

2.5 Toward an aligned future

'When forced to work within a strict framework, the imagination is taxed to its utmost–and will produce its richest ideas. Given total freedom the work is likely to sprawl.'
TS Eliot

'We can love the beautiful, and believe in it, and thereby open ourselves to an understanding of love that does not dominate, but cherishes the independence and beauty of the loved.'
Martha Nussbaum

What does the opposite of overcommitment look like in a future co-created by AI? In living systems, evolving semi-permeable boundaries contextualize partial forms to be more long-sighted in time and space, increasing subtlety and potential. Now we take a preliminary look at applying what we learn from life to create an aligned future.

The only way to be truly aligned, is to be aligned with all the subtlety of the world and all the potential of it, which isn't captured by any formalism.

Good outcomes will never happen if commitments are too strong. Is this just tautologically saying that 'too much is too much'? The central message of this paper is that it's very interesting that too much anything is bad. This does not just mean 'too many bags of carrots' but

includes excess emphasis on literally any pattern – a definition, a conceptual framework, even homogeneity (cites). What stops a pattern from becoming 'too much'? The web of constraints it's nestled into. Whether they're slowing down a rampaging conqueror or a homogenizing diffusion, the boundaries are somehow holding in check a particular thing from becoming 'too much'.

The fully general version of this is not well integrated into AI safety. It's integrated in special cases, like the ones we highlight: midas problem, inequality, conceptual monoculture. Those are all cases where hyperexaggerating one pattern leads to bad futures. But the general version means that excess attachment to literally anything (an architectural motif, a way of doing training, a conceptual framework, a definition, a kind of social patterning) is bad. This is so relevant to AI safety because AI is exactly the technology that takes particular forms and runs away with them.

Alignment can be viewed as the problem of avoiding overcommitment to any particular form. No matter how good or complete our current concepts or specifications or lenses are, treating them as absolutes is not aligned. A system is misaligned if it has a cap on its sensitivity to a larger context.

Finding a balance between existing factors is necessary but not sufficient. Alignment means staying at the subtle edge of sensitivity where fundamentally new things are entering. As humans, there's always *something* we're taking as axiomatically true – something that's part of the structure of us as the thinker. But if we're right at that edge, we sometimes step back to see the axiomatic thing in context. There's always more to be sensitive to, something outside the available modalities of the thinker's perception. The stepping back process inherently can't be fully conceptualized. But it's what life does, that's the creative force of life that rides the knife edge. And that's what real alignment has to do (for the system including AI and humans). Fixing alignment to something conceptualizable would be disaster if AI becomes powerful.

2.5.1 Semi-permeable boundaries and contextualization

The use of boundaries against overcommitment in technology is as old as technology. As soon as we started building things, we had to build in boundaries, because we want the things we build to be robust and useful and not to collapse into degenerate states. We put circuit breakers in the power grid, governors in steam engines, escapements in clocks, ReLUs in neural networks. In modern AI research we have personalization through the context, access to user data, on-device learning. We have conditional computation creating separation between parts within a large model. We have dropout, cross-validation, causal masking, multi-agent systems. Many safety methods are boundaries, including safety post-training, guardrail models, red teaming, mechanistic interpretability, government oversight and so on (Gabriel et al., 2025). Increasingly critical are ongoing evaluation and monitoring of deployed AI systems (Grey and Segerie, 2025; Myllyaho et al., 2021; Yampolskiy, 2025). There is increasing awareness that over-attachment to fixed optimization targets is often counterproductive (Kumar et al., 2025; Stanley and Lehman, 2015; Stanley et al., 2017).

To mitigate this, the field has moved toward solutions:

- *Mechanistic interpretability.* Improving our mechanistic understanding of AI systems

so we can, for example, detect and correct the systems if they develop hidden ways of resisting our efforts to change their goals (Anthropic Research Team, 2024; Bereska and Gavves, 2024; Burns et al., 2022; Olah et al., 2020).

- *Deliberated preferences or idealized values.* Representing what we would prefer if we had more time, knowledge, and computational power (Bostrom, 2014; Soares and Fallenstein, 2014; Yudkowsky, 2004). One way to access longer-sighted preferences is by giving humans more time and resources to think about their answer, to ask on behalf of another person, or on behalf of their future self. You can give them access to tools and information. You can ask people retrospectively whether an outcome was good, rather than prospectively.

- *Pluralistic alignment.* (Sorensen et al., 2024). Tries to do X. Limitations of pluralistic alignment. There is not even universal agreement on which principles are most appropriate for aggregating the preferences of different people. Pluralistic alignment faces the risk of overcommitment if it treats the aggregated values or agreed principles (or even the mechanism of aggregation or discovery) as absolute. If we formalize a democratic process and maximize adherence to it, we risk tyranny of the majority or the entrenchment of biases codified in the aggregation algorithm (Gabriel, 2020).

- Principles frameworks.

- Inverse methods to learn values.

- Methods for limiting the power or capabilities of AI systems. Conservatism (Cohen et al., 2024). Penalize large or irreversible changes to the environment, regardless of apparent reward gains. Adds an auxiliary term or constraint that discourages high-impact actions unless clearly beneficial, thereby making overoptimization costly. Attainable Utility Preservation (Turner et al., 2020). Relative reachability and impact measures (Krakovna et al., 2018). Conservative agency. Myopic RL and 'one-shot' decision frameworks (Hubinger et al. 2021). Quantilization (Everitt et al. 2017) – sample from a safe baseline rather than fully optimize.

Another important method in the field is open-endedness. For example, novelty search. Can briefly summarize Stanley et al here. Of course, even any particular way of approaching this will have the same issues.

At a social level, we have labs taking different approaches, nations with different cultures and strategic interests, ideas drawn from diverse fields like neuroscience and physics.

How can we use AI in a way that does not lead to domination of individual perspectives? (and we might want to distinguish 'concentration of power in the hands of elites' versus 'collapse of global thought diversity without anyone necessarily benefitting'). Policies and regulations. Education, making sure everyone gets good at utilizing what is available.

Setting up the structures that support this continued expansion of internal space. Generation of new forms, moving away from equilibrium.

Iterated amplification, where AI helps us think better, and we iterate on that cycle.

Pluralistic alignment (Sorensen et al., 2024). Processes for deliberation and inclusion.

Principles that prevent over-reach of any one party (Gabriel and Keeling, 2025). 'when a technology has profound societal effects it ought to be regulated by principles that are amenable to public rather than private justification'... 'efforts to align AI systems with a given moral schema may lead to unjust value imposition or even domination'

One perhaps underexplored question is how to preserve and enhance diversity in human culture and concepts, at scales anywhere from national or ethnic groups to individuals, while productively putting the diverse elements in contact with each other. In recent decades, information exchange has enormously increased with telecommunications, the internet, and now AI itself. Because we can now use AI to effectively get answers from the rest of humanity instantly, an important aspect of this question is how we will structure our own use of AI to maintain our autonomy and diversity. (can refer back to the collapses described in 'problem solving in groups' and 'conceptual monoculture')

This proliferation of structure is, under the framework of this paper, potentially extremely valuable and aligned. The danger is if it falls off the knife edge and collapses into homogeneity, randomness or rigidity.

How can we use AI systems in a more thoughtful way that could increase rather than decrease multiscale nuance and livingness?

Avoiding overcommitment means maintaining sensitivity to the actual larger context of the world. It's almost infinitely nuanced.

So, alignment is sensitivity to the real context of existing form. This is why it matches pretty well to existing definitions of alignment and human values: because these things already exist and are part of the world. Hopefully it can also go beyond them. It doesn't make sense to lock-in to present human values and concepts (cite Bostrom).

2.5.2 An ongoing process

But we hope our perspective also hints at something more. Even the most foundational assumptions are probably not final answers. The lenses we use to look at the world keep changing. True alignment is a process where whatever was previously axiomatic becomes a contextualized object. The point of alignment is not to say that any particular perspective is absolutely wrong or right. An aligned future will include continual reinvention of whatever concepts we have, including to the assumptions those concepts are built on, and the assumptions those assumptions are built on. Whatever concepts we currently have do not place hard limits on the future. Even the concept of 'not being overly attached to our concepts' is itself something we can release into contextualization.

As humans – whether AI researchers or any participants in social systems – this can be a mundane practical process of holding ideas with some skepticism, having patience to look at different timescales, listening to an internal voice of wisdom, entertaining conflicting perspectives. It can also be a profound process of self-awareness and personal growth. We continually evolve what we believe, even our self-definition. We become sensitive to more and more aspects of the extaordinarily subtle world, releasing beliefs into larger awareness without losing or erasing them.

Perhaps the most interesting question is this: what does the process of open-ended contextualization look like within AI systems, or in human-AI relationships? Is there a version of AI that continually contextualizes its own processes as partial truths? Do existing AI systems already do this to some degree, as they train and as they learn from interactions with humans and the world? What would it mean to take this process farther, for AI to continually release from exclusive attachment to any particular form? What can we do now to protect the potential for even any form of that releasing process to not be a final answer?

If we attach too strongly to any conceptualization of what we mean by 'avoiding overcommitment' or 'life-like evolution', then an AI system could masquarade as life-like without truly being so.

2.6   Conclusion

We learn from living systems that health and flourishing is not any particular form or concept. Therefore, alignment is not picking the right values or principles, or even the right system for learning them. It is not any method for interpretability or corrigibility. All of these can be useful parts of alignment. But alignment itself is the continued dance of contextualizing any particular form, more fully respecting the nearly unfathomable subtlety of the existing world, and open-endedly supporting new depth that admits more and more different kinds of metaphor or description. In this way, AI can participate in intense flourishing of an evolving world even far beyond current human conceptualization.

Acknowledgements

Competing Interests

The authors declare no competing interests.

References

R. A. Adams, K. E. Stephan, H. R. Brown, C. D. Frith, and K. J. Friston. The computational anatomy of psychosis. *Frontiers in psychiatry*, 4:47, 2013.

T. W. Adorno, E. Frenkel-Brunswik, D. J. Levinson, and R. N. Sanford. *The Authoritarian Personality*. Harper & Brothers, New York, 1950.

D. Agarwal, M. Naaman, and A. Vashistha. Ai suggestions homogenize writing toward western styles and diminish cultural nuances. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2025.

G. Ainslie. *Picoeconomics: The Strategic Interaction of Successive Motivational States within the Person.* Cambridge University Press, Cambridge, 1992.

M. D. S. Ainsworth, M. C. Blehar, E. Waters, and S. Wall. *Patterns of attachment: A psychological study of the strange situation.* Lawrence Erlbaum Associates, Hillsdale, NJ, 1978.

B. Alberts, R. Heald, A. Johnson, D. Morgan, M. Raff, K. Roberts, and P. Walter. *Molecular biology of the cell: seventh international student edition with registration card.* WW Norton & Company, 2022.

S. Ambec, M. A. Cohen, S. Elgie, and P. Lanoie. The porter hypothesis at 20: can environmental regulation enhance innovation and competitiveness? *Review of environmental economics and policy*, 7(1):2–28, 2013.

D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

L. R. Anderson and C. A. Holt. Information cascades in the laboratory. *The American economic review*, pages 847–862, 1997.

P. W. Anderson. More is different: Broken symmetry and the nature of the hierarchical structure of science. *Science*, 177(4047):393–396, Aug 1972. doi: 10.1126/science.177.4047.393.

Anthropic Research Team. Mapping the mind of a large language model. Online research report / blog post, 2024.

APA. *Diagnostic and statistical manual of mental disorders.* American psychiatric association, 2013.

N. A. Ashford, C. Ayers, and R. F. Stone. Using regulation to change the market for innovation. *Harv. Envtl. L. Rev.*, 9:419, 1985.

P. Bak. *How nature works: the science of self-organized criticality.* Springer Science & Business Media, 2013.

P. Bak, C. Tang, and K. Wiesenfeld. Self-organized criticality: An explanation of 1/f noise. *Physical Review Letters*, 59:381–384, 1987. doi: 10.1103/PhysRevLett.59.381.

J. Bakan. The corporation. the pathological pursuit of profit and power, 2006.

T. Ballard, J. B. Vancouver, and A. Neal. On the pursuit of multiple goals with different deadlines. *Journal of Applied Psychology*, 103(11):1242, 2018.

B. W. Balleine, M. R. Delgado, and O. Hikosaka. The role of the dorsal striatum in reward and decision-making. *Journal of Neuroscience*, 27(31):8161–8165, 2007.

P. A. Baran. *Monopoly capital.* NYU Press, 1966.

R. S. Baron. So right it's wrong: Groupthink and the ubiquitous nature of polarized group decision making. *Advances in experimental social psychology*, 37(2):219–253, 2005.

J. A. Baross and S. E. Hoffman. Submarine hydrothermal vents and associated gradient environments as sites for the origin and evolution of life. *Origins of Life and Evolution of the Biosphere*, 15(4):327–345, 1985.

G. Bateson. *Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology*. Chandler Publishing Company, San Francisco, 1972. ISBN 0810204479.

S. D. Baum. Social choice ethics in artificial intelligence. *Ai & Society*, 35(1):165–176, 2020.

S. Bazazi, J. von Zimmermann, B. Bahrami, and D. Richardson. Self-serving incentives impair collective decisions by increasing conformity. *PloS one*, 14(11):e0224725, 2019.

L. A. Bebchuk and A. Cohen. The costs of entrenched boards. *Journal of financial economics*, 78(2):409–433, 2005.

J. S. Beck, A. Beck, and J. Beck. Cognitive behavior therapy: basics and beyond. ed. *New York*, 2011.

J. Becker, D. Brackbill, and D. Centola. Network dynamics of social influence in the wisdom of crowds. *Proceedings of the national academy of sciences*, 114(26):E5070–E5076, 2017.

J. M. Beggs and D. Plenz. Neuronal avalanches in neocortical circuits. *Journal of neuroscience*, 23(35):11167–11177, 2003.

N. Beguš. Experimental narratives: A comparison of human crowdsourced storytelling and ai storytelling. *Humanities and Social Sciences Communications*, 11(1):1–22, 2024.

T. E. Behrens, T. H. Muller, J. C. Whittington, S. Mark, A. B. Baram, K. L. Stachenfeld, and Z. Kurth-Nelson. What is a cognitive map? organizing knowledge for flexible behavior. *Neuron*, 100(2):490–509, 2018.

A. Beisser. The paradoxical theory of change. *Gestalt Therapy Now*, 1970. Reference for Gestalt approach to awareness and change.

G. Bell. *The Masterpiece of Nature: The Evolution and Genetics of Sexuality*. University of California Press, Berkeley, 1982. ISBN 978-0520045835.

L. Bereska and E. Gavves. Mechanistic interpretability for ai safety–a review. *arXiv preprint arXiv:2404.14082*, 2024.

I. Berlin. Two concepts of liberty. In I. Berlin, editor, *Four Essays on Liberty*, pages 118–172. Oxford University Press, Oxford, 1969. ISBN 9780192810340.

E. Bernstein, J. Shore, and D. Lazer. How intermittent breaks in interaction improve collective intelligence. *Proceedings of the National Academy of Sciences*, 115(35):8734–8739, 2018.

A. Birhane, W. Isaac, V. Prabhakaran, M. Diaz, M. C. Elish, I. Gabriel, and S. Mohamed. Power to the people? opportunities and challenges for participatory ai. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–8, 2022.

A. Bobu, A. Bajcsy, J. F. Fisac, S. Deglurkar, and A. D. Dragan. Quantifying hypothesis space misspecification in learning from human–robot demonstrations and physical corrections. *IEEE Transactions on Robotics*, 36(3):835–854, 2020.

B. Bodhi. *The Connected Discourses of the Buddha: A Translation of the Samyutta Nikaya*. Wisdom Publications, Somerville, MA, 2000. Reference for the Sallatha Sutta (The Arrow).

R. Bommasani. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

N. Bostrom. *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press, 2014. ISBN 0199678111.

M. M. Botvinick, T. S. Braver, D. M. Barch, C. S. Carter, and J. D. Cohen. Conflict monitoring and cognitive control. *Psychological review*, 108(3):624, 2001.

M. Bowen. *Family therapy in clinical practice.* Jason Aronson, 1978.

G. E. P. Box. Science and statistics. *Journal of the American Statistical Association*, 71(356): 791–799, 1976.

J. Bradshaw. *Healing the shame that binds you.* Health Communications, Inc., 1988.

R. B. Brandt. *A Theory of the Good and the Right.* Clarendon Press, Oxford, 1979.

T. S. Braver. The variable nature of cognitive control: a dual mechanisms framework. *Trends in cognitive sciences*, 16(2):106–113, 2012.

D. Bray. *Wetware: a computer in every living cell.* Yale University Press, 2019.

R. Bregman. *Humankind: A Hopeful History.* Bloomsbury Publishing, London, 2020. ISBN 9781408898956.

B. Brown. *Daring Greatly: How the Courage to Be Vulnerable Transforms the Way We Live, Love, Parent, and Lead.* Gotham Books, New York, NY, 2012. ISBN 9781592407330.

P. Brown. A rhythmic mechanism for communication in the cortex. *Trends in neurosciences*, 26(5):232–233, 2003.

E. Brynjolfsson, D. Li, and L. Raymond. Generative ai at work. *The Quarterly Journal of Economics*, 140(2):889–942, 2025.

C. Burns, H. Ye, D. Klein, and J. Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

N. Burns and S. Kedia. The impact of performance-based compensation on misreporting. *Journal of financial economics*, 79(1):35–67, 2006.

D. T. Campbell. Assessing the impact of planned social change. *Evaluation and program planning*, 2(1):67–90, 1979.

G. Canguilhem. *Le normal et le pathologique.* Presses Universitaires de France, Paris, 1966.

R. N. Cardinal, J. A. Parkinson, J. Hall, and B. J. Everitt. Emotion and motivation: the role of the amygdala, ventral striatum, and prefrontal cortex. *Neuroscience & Biobehavioral Reviews*, 26(3):321–352, 2002.

P. Carnes. *Out of the shadows: Understanding sexual addiction.* Hazelden Publishing, 2001.

CCIA Research Center. 2025 survey of product impact in the connected economy: Artificial intelligence. Spice ai report, Computer & Communications Industry Association, Nov. 2025. URL https://ccianet.org/research/reports/

2025-survey-of-product-impact-in-the-connected-economy-artificial-intelligence/. Accessed: 2025-12-08.

G. Ceballos, P. R. Ehrlich, A. D. Barnosky, A. García, R. M. Pringle, and T. M. Palmer. Accelerated modern human–induced species losses: Entering the sixth mass extinction. *Science advances*, 1(5):e1400253, 2015.

A. J. Chaney, B. M. Stewart, and B. E. Engelhardt. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM conference on recommender systems*, pages 224–232, 2018.

B. Charlesworth, M. T. Morgan, and D. Charlesworth. The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134(4):1289–1303, 1993.

A. Chatterji, T. Cunningham, D. Deming, Z. Hitzig, C. Ong, C. Shan, and K. Wadman. How people use ChatGPT. Technical report, OpenAI, Sept. 2025. URL https://cdn.openai.com/pdf/a253471f-8260-40c6-a2cc-aa93fe9f142e/economic-research-chatgpt-usage-paper.pdf.

D. R. Chialvo. Emergent complex neural dynamics. *Nature physics*, 6(10):744–750, 2010.

N. Chomsky. *Syntactic Structures*. Mouton de Gruyter, The Hague, 1957.

A. Clark. *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford University Press, New York, NY, USA, 2008. ISBN 9780195374443.

H. Cloud and J. Townsend. *Boundaries: When to Say Yes, How to Say No to Take Control of Your Life*. Zondervan, Grand Rapids, MI, 1992.

J. Clune, J.-B. Mouret, and H. Lipson. The evolutionary origins of modularity. *Proceedings of the Royal Society b: Biological sciences*, 280(1755):20122863, 2013.

L. Cocchi, L. L. Gollo, A. Zalesky, and M. Breakspear. Criticality in the brain: A synthesis of neurobiology, models and cognition. *Progress in neurobiology*, 158:132–152, 2017.

M. K. Cohen, N. Kolt, Y. Bengio, G. K. Hadfield, and S. Russell. Regulating advanced artificial agents. *Science*, 384(6691):36–38, 2024.

L. L. Colgin, T. Denninger, M. Fyhn, T. Hafting, T. Bonnevie, O. Jensen, M.-B. Moser, and E. I. Moser. Frequency of gamma oscillations routes flow of information in the hippocampus. *Nature*, 455:125–129, 2008. doi: 10.1038/nature07278.

M. Condorcet. *Essai sur l'Application de l'Analyse a la Probabilité des Décisions Rendues a la Pluralité des Voix*. Imprimerie Royale, Paris, 1785.

T. H. Costello, G. Pennycook, and D. G. Rand. Durably reducing conspiracy beliefs through dialogues with ai. *Science*, 2024.

J. P. Craven. *The Silent War: The Cold War Battle Beneath the Sea*. Simon and Schuster, 2002.

K. Crawford. *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press, 2021.

J. F. Crow and M. Kimura. Evolution in sexual and asexual populations. *The American Naturalist*, 99(909):439–450, 1965. doi: 10.1086/282389.

J. P. Crutchfield. The calculi of emergence: computation, dynamics and induction. *Physica D: Nonlinear Phenomena*, 75(1-3):11–54, 1994.

A. Dafoe. Ai governance: a research agenda. *Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK*, 1442:1443, 2018.

W. Dalrymple. *The Anarchy: The Relentless Rise of the East India Company*. Bloomsbury Publishing, 2019. ISBN 9781408864401. URL `https://books.google.co.uk/books?id=-T2lDwAAQBAJ`.

R. Dawkins. *The Selfish Gene*. Oxford University Press, Oxford, 1976.

P. Dayan, Y. Niv, B. Seymour, and N. D. Daw. The misbehavior of value and the discipline of the will. *Neural networks*, 19(8):1153–1160, 2006.

E. De Bono. Lateral thinking. *New York*, page 70, 1970.

G. Deco and V. K. Jirsa. Ongoing cortical activity at rest: criticality, multistability, and ghost attractors. *Journal of Neuroscience*, 32(10):3366–3375, 2012.

G. Deco, V. K. Jirsa, and A. R. McIntosh. Emerging concepts for the dynamical organization of resting-state activity in the brain. *Nature Reviews Neuroscience*, 12(1):43–56, 2011.

S. Dehaene, F. Al Roumi, Y. Lakretz, S. Planton, and M. Sablé-Meyer. Symbols and mental programs: a hypothesis about human singularity. *Trends in Cognitive Sciences*, 26(9):751–766, 2022.

J. Dewey. Theory of valuation. *International encyclopedia of unified science*, 1939.

M. Diehl and W. Stroebe. Productivity loss in brainstorming groups: Toward the solution of a riddle. *Journal of personality and social psychology*, 53(3):497, 1987.

S. Dohnány, Z. Kurth-Nelson, E. Spens, L. Luettgau, A. Reid, C. Summerfield, M. Shanahan, and M. M. Nour. Technological folie a deux: Feedback loops between ai chatbots and mental illness. *arXiv preprint arXiv:2507.19218*, 2025.

A. R. Doshi and O. P. Hauser. Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science advances*, 10(28):eadn5290, 2024.

D. G. Douglas. *The social construction of technological systems, anniversary edition: New directions in the sociology and history of technology*. MIT press, 2012.

R. J. Douglas and K. A. Martin. Neuronal circuits of the neocortex. *Annu. Rev. Neurosci.*, 27(1):419–451, 2004.

L. Drago and R. Laine. Defining the intelligence curse. `https://intelligence-curse.ai/defining/`, April 2025. Accessed: 2025-11-05.

H. L. Dreyfus. *What Computers Can't Do: The Limits of Artificial Intelligence*. Harper & Row, New York, NY, 1972.

K. Duncker. On problem-solving. *Psychological Monographs*, 58, 1945.

J. L. Evenden. Varieties of impulsivity. *Psychopharmacology*, 146(4):348–361, 1999.

P. Federn. Narcissism in the structure of the ego. *The International Journal of Psycho-Analysis*, 9:401, 1928.

D. C. Feldman. Who's socializing whom? the impact of socializing newcomers on insiders, work groups, and organizations. *Human Resource Management Review*, 4(3):213–233, 1994.

J. Felsenstein. The evolutionary advantage of recombination. *Genetics*, 78(2):737–756, 1974. doi: 10.1093/genetics/78.2.737.

Y.-J. Feng, D. C. Blackburn, D. Liang, D. M. Hillis, D. B. Wake, D. C. Cannatella, and P. Zhang. Phylogenomics reveals rapid, simultaneous diversification of three major clades of gondwanan frogs at the cretaceous–paleogene boundary. *Proceedings of the national Academy of Sciences*, 114(29):E5864–E5870, 2017.

P. K. Feyerabend. *Against Method*. Verso, 1975.

R. Firth. Ethical absolutism and the ideal observer. *Philosophy and Phenomenological Research*, 12(3):317–345, 1952.

R. A. Fisher. *The Genetical Theory of Natural Selection*. The Clarendon Press, Oxford, 1930.

M. L. Flowers. A laboratory test of some implications of janis's groupthink hypothesis. *Journal of Personality and Social Psychology*, 35(12):888, 1977.

J. A. Fodor. *The Language of Thought*. Harvard University Press, Cambridge, MA, 1975.

C. Folke. Resilience: The emergence of a perspective for social–ecological systems analyses. *Global environmental change*, 16(3):253–267, 2006.

M. Ford. *Rise of the Robots: Technology and the Threat of a Jobless Future*. Basic Books, New York, 2015.

A. Frank. To find alien intelligence, start with the mountains. *The Atlantic*, November 2024. URL https://www.theatlantic.com/science/archive/2024/11/cambrian-explosion-tectonic-plates-mountains-evolution/680544/. Accessed December 25, 2024.

M. J. Frank and E. D. Claus. Anatomy of a decision: striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. *Psychological review*, 113(2):300, 2006.

A. Freud. *Das Ich und die Abwehrmechanismen*. Internationaler Psychoanalytischer Verlag, Wien, 1936.

S. Freud. The neuro-psychoses of defence. *Collected Papers*, 3:45–61, 1894. Originally published as: Die Abwehr-Neuropsychosen, 1894, Neurologisches Centralblatt, 13, 4, 50–51, 54–61.

S. Freud. The future prospects of psycho-analytic therapy. *Collected Papers*, 2:285–296, 1910. Originally published as: Über die zukünftigen Chancen der psychoanalytischen Therapie, 1910, Zentralblatt für Psychoanalyse, 1, 7, 297–311.

S. Freud. *Totem und Tabu: Einige Übereinstimmungen im Seelenleben der Wilden und der Neurotiker*. Hugo Heller & Cie, Leipzig und Wien, 1913.

V. Frey and A. Van de Rijt. Social influence undermines the wisdom of the crowd in sequential decision making. *Management science*, 67(7):4273–4286, 2021.

G. O. Gabbard and E. P. Lester. *Boundaries and boundary violations in psychoanalysis*. American Psychiatric Publishing, 1995.

I. Gabriel. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437, 2020.

I. Gabriel and G. Keeling. A matter of principle? ai alignment as the fair treatment of claims. *Philosophical Studies*, pages 1–23, 2025.

I. Gabriel, G. Keeling, A. Manzini, and J. Evans. We need a new ethics for a world of AI agents. *Nature*, 644(8075):38–40, Aug. 2025. doi: 10.1038/d41586--025-02454--5. URL https://www.nature.com/articles/d41586--025-02454--5. Comment.

H.-G. Gadamer. *Wahrheit und Methode*. J.C.B. Mohr (Paul Siebeck), 1960.

D. D. Garrett, G. R. Samanez-Larkin, S. W. MacDonald, U. Lindenberger, A. R. McIntosh, and C. L. Grady. The bold brain: greater variability of bold t2* signal is associated with better cognitive performance. *Journal of Neuroscience*, 33(2):835–840, 2013.

R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

M. Gell-Mann. *The Quark and the Jaguar: Adventures in the Simple and the Complex*. Macmillan, 1994.

G. Gigerenzer and W. Gaissmaier. Heuristic decision making. *Annual review of psychology*, 62:451–482, 2011.

T. Gilovich, D. Griffin, and D. Kahneman, editors. *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge University Press, Cambridge, 2002. ISBN 978-0521796798.

E. Goffman. *Frame analysis: An essay on the organization of experience*. Harvard university press, 1974.

B. Goldacre. *Bad pharma: how drug companies mislead doctors and harm patients*. Macmillan, 2014.

N. Goldenfeld and C. Woese. Life is physics: evolution as a collective phenomenon far from equilibrium. *Annu. Rev. Condens. Matter Phys.*, 2(1):375–399, 2011.

U. Goodenough and J. Heitman. Origins of eukaryotic sexual reproduction. *Cold Spring Harbor perspectives in biology*, 6(3):a016154, 2014.

N. Goodman. *Ways of worldmaking*, volume 51. Hackett Publishing, 1978.

I. Gough. Universal basic services: A theoretical and moral framework. *The Political Quarterly*, 90(3):534–542, 2019.

C. L. Grady and D. D. Garrett. Understanding variability in the bold signal and why it matters for aging. *Brain Imaging and Behavior*, 8:274–282, 2014. doi: 10.1007/s11682--013-9253--0.

J. Greene. *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them.* The Penguin Press, New York, NY, 2013. ISBN 9781594202605.

M. Grey and C.-R. Segerie. Safety by measurement: a systematic literature review of ai safety evaluation methods. *arXiv preprint arXiv:2505.05541*, 2025.

S. J. Grossman and O. D. Hart. The costs and benefits of ownership: A theory of vertical and lateral integration. *Journal of political economy*, 94(4):691–719, 1986.

K. Hackenburg, B. M. Tappin, L. Hewitt, E. Saunders, S. Black, H. Lin, C. Fist, H. Margetts, D. G. Rand, and C. Summerfield. The levers of political persuasion with conversational artificial intelligence. *Science*, 390(6777):eaea3884, 2025.

D. Hadfield-Menell and G. K. Hadfield. Incomplete contracting and ai alignment. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 417–422, 2019.

D. Hadfield-Menell, A. D. Dragan, P. Fisac, and S. Russell. Cooperative inverse reinforcement learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, pages 3909–3917, 2016.

D. Hadfield-Menell, A. D. Dragan, and S. Russell. The off-switch game. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI 2017)*, pages 220–227, 2017.

J. Haidt. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4):814–834, 2001. doi: 10.1037/0033-295X.108.4.814.

J. Haidt. *The righteous mind: Why good people are divided by politics and religion.* Vintage, 2012.

A. G. Haldane. Rethinking the financial network. In *Fragile stabilität–stabile fragilität*, pages 243–278. Springer, 2013.

C. Haldeman and J. M. Beggs. Critical branching captures activity in living neural networks and maximizes the number of metastable states. *Physical Review Letters*, 94(5):058101, 2005. doi: 10.1103/PhysRevLett.94.058101.

C. Hammond, H. Bergman, and P. Brown. Pathological synchronization in parkinson's disease: networks, models and treatments. *Trends in neurosciences*, 30(7):357–364, 2007.

D. J. Haraway. *Staying with the trouble: Making kin in the Chthulucene.* Duke University Press, 2020.

F. M. Harold. *The way of the cell: molecules, organisms, and the order of life.* Oxford University Press, 2001.

H. Hart. Harmony in irony, jul 2025. URL https://stemless.substack.com/p/harmony-in-irony. Substack post.

H. K. Hartline, H. G. Wagner, and F. Ratliff. Inhibition in the eye of limulus. *The Journal of general physiology*, 39(5):651–673, 1956.

G. Hatano and K. Inagaki. Two courses of expertise. *Clinical Center for Early Childhood Development Annual Report*, 6:27–36, 1984.

K. E. Hauer, L. Edgar, S. O. Hogan, B. Kinnear, and E. Warm. The science of effective group process: lessons for clinical competency committees. *Journal of Graduate Medical Education*, 13(2 Suppl):59, 2021.

P. H. Hawley. Prosocial and coercive configurations of resource control in early adolescence: A case for the well-adapted machiavellian. *Merrill-Palmer Quarterly*, 49(3):279–309, 2003.

N. K. Hayles. How we became posthuman: Virtual bodies in cybernetics, literature, and informatics, 2000.

R. M. Hazen and D. A. Sverjensky. Mineral surfaces, geochemical complexities, and the origins of life. *Cold Spring Harbor perspectives in biology*, 2(5):a002162, 2010.

M. Heidegger. Letter on humanism. In W. McNeill, editor, *Pathmarks*. Cambridge University Press, Cambridge, 1998. Originally written 1946.

A. Heinz, G. K. Murray, F. Schlagenhauf, P. Sterzer, A. A. Grace, and J. A. Waltz. Towards a unifying cognitive, neurophysiological, and computational neuroscience account of schizophrenia. *Schizophrenia bulletin*, 45(5):1092–1100, 2019.

W. A. Hershberger. An approach through the looking-glass. *Animal Learning & Behavior*, 14 (4):443–451, 1986.

S. M. Herzog and R. Hertwig. Harnessing the wisdom of the inner crowd. *Trends in cognitive sciences*, 18(10):504–506, 2014.

W. G. Hill and A. Robertson. The effect of linkage on limits to artificial selection. *Genetical Research*, 8(3):269–294, 1966. PMID: 5980116.

D. R. Hofstadter. Analogy as the core of cognition. *The analogical mind: Perspectives from cognitive science*, pages 499–538, 2001.

R. M. Hogarth. A note on aggregating opinions. *Organizational behavior and human performance*, 21(1):40–46, 1978.

J. H. Holland. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. University of Michigan Press, Ann Arbor, MI, 1975. ISBN 0472084607.

C. S. Holling. Resilience and stability of ecological systems. *Annual Review of Ecology and Systematics*, 4:1–23, 1973.

Honeywell. Honeywell and Google Cloud to accelerate autonomous operations with AI agents for the industrial sector, Oct. 2024. URL https://www.honeywell.com/us/en/press/2024/10/honeywell-and-google-cloud-to-accelerate-autonomous-operations-with-ai-agents-for-the. Press Release.

L. Hong and S. E. Page. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101(46):16385–16389, 2004.

D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106, 1962.

R. R. Hudson and N. L. Kaplan. Deleterious background selection with recombination. *Genetics*, 141(4):1605–1617, 1995.

P. Humphreys. *Extending Ourselves: Computational Science, Empiricism, and Scientific Method.* Oxford University Press, New York, NY, USA, 2004. ISBN 9780195158700.

A. A. Hung and C. R. Plott. Information cascades: Replication and an extension to majority rule and conformity-rewarding institutions. *American Economic Review*, 91(5):1508–1520, 2001.

G. E. Hutchinson. Homage to santa rosalia or why are there so many kinds of animals? *The American Naturalist*, 93(870):145–159, 1959.

J. S. Isaacson and M. Scanziani. How inhibition shapes cortical activity. *Neuron*, 72(2):231–243, 2011.

D. Jablonski. Mass extinctions and macroevolution. *Paleobiology*, 31(S2):192–210, 2005.

I. L. Janis. *Victims of groupthink: A psychological study of foreign-policy decisions and fiascoes.* Houghton Mifflin, 1972.

K. Jaspers. *General psychopathology*, volume 2. JHU Press, 1997.

K. Javed and R. S. Sutton. The big world hypothesis and its ramifications for artificial intelligence. In *Finding the Frame: An RLC Workshop for Examining Conceptual Frameworks*, 2024.

O. Jensen and A. Mazaheri. Shaping functional architecture by oscillatory alpha activity: gating by inhibition. *Frontiers in human neuroscience*, 4:186, 2010.

H. J. Jeon, S. Milli, and A. Dragan. Reward-rational (implicit) choice: A unifying formalism for reward learning. *Advances in Neural Information Processing Systems*, 33:4415–4426, 2020.

Y. J. John, L. Caldwell, D. E. McCoy, and O. Braganza. Dead rats, dopamine, performance metrics, and peacock tails: Proxy failure is an inherent risk in goal-oriented systems. *Behavioral and Brain Sciences*, pages 1–68, 2023.

C. G. Jung. *The Archetypes and the Collective Unconscious (Collected Works of C.G. Jung, Vol. 9, Part 1).* Princeton University Press, Princeton, NJ, 1969. Reference for The Shadow and resistance mechanisms.

D. Kahneman. *Thinking, Fast and Slow.* Farrar, Straus and Giroux, New York, 2011. ISBN 978-0374275631.

A. A. Kane, L. Argote, and J. M. Levine. Knowledge transfer between groups via personnel rotation: Effects of social identity and knowledge quality. *Organizational behavior and human decision processes*, 96(1):56–71, 2005.

T. B. Kashdan and J. Rottenberg. Psychological flexibility as a fundamental aspect of health. *Clinical psychology review*, 30(7):865–878, 2010.

S. A. Kauffman. The origins of order: Self-organization and selection in evolution. In *Spin glasses and biology*, pages 61–100. World Scientific, 1992.

P. D. Keightley and S. P. Otto. Interference among deleterious mutations favours sex and recombination in finite populations. *Nature*, 443(7107):89–92, 2006.

S. Kerr. On the folly of rewarding a, while hoping for b. *Academy of Management journal*, 18 (4):769–783, 1975.

M. Kirchhoff, T. Parr, E. Palacios, K. Friston, and J. Kiverstein. The markov blankets of life: autonomy, active inference and the free energy principle. *Journal of The royal society interface*, 15(138):20170792, 2018.

H. R. Kirk, I. Gabriel, C. Summerfield, B. Vidgen, and S. A. Hale. Why human–ai relationships need socioaffective alignment. *Humanities and Social Sciences Communications*, 12(1):1–9, 2025.

R. Kirk, I. Mediratta, C. Nalmpantis, J. Luketina, E. Hambro, E. Grefenstette, and R. Raileanu. Understanding the effects of rlhf on llm generalisation and diversity. *arXiv preprint arXiv:2310.06452*, 2023.

J. Kleinberg and M. Raghavan. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118, 2021.

W. Klimesch, P. Sauseng, and S. Hanslmayr. Eeg alpha oscillations: the inhibition–timing hypothesis. *Brain research reviews*, 53(1):63–88, 2007.

E. Kolbert. *The sixth extinction: An unnatural history.* Henry Holt and Company, 2014.

A. Korzybski. *Science and Sanity: An Introduction to Non-Aristotelian Systems and General Semantics.* The International Non-Aristotelian Library Publishing Company, Lancaster, PA, 1933.

N. Kosmyna, E. Hauptmann, Y. T. Yuan, J. Situ, X.-H. Liao, A. V. Beresnitzky, I. Braunstein, and P. Maes. Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing task. *arXiv preprint arXiv:2506.08872*, 2025.

S. Kotler, M. Mannino, K. Friston, G. Buzsáki, J. S. Kelso, and G. Dumas. Pathfinding: a neurodynamical account of intuition. *Communications Biology*, 8(1):1214, 2025.

V. Krakovna, L. Orseau, R. Kumar, M. Martic, and S. Legg. Penalizing side effects using stepwise relative reachability. *arXiv preprint arXiv:1806.01186*, 2018.

V. Krakovna, A. Gleave, and J. Miller. Specification gaming: The flip side of AI ingenuity. DeepMind Safety Research Blog, May 2020. URL `https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity`. Accessed on 2025–10-09.

S. W. Kraus, V. Voon, and M. N. Potenza. Should compulsive sexual behavior be considered an addiction? *Addiction*, 111(12):2097–2106, 2016.

J. Krishnamurti. *Freedom from the Known.* Harper & Row, New York, NY, 1969. Reference for 'The Observer is the Observed'.

T. S. Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago, 2nd edition, 1970. ISBN 9780226458083.

A. Kumar, J. Clune, J. Lehman, and K. O. Stanley. Questioning representational optimism in deep learning: The fractured entangled representation hypothesis. *arXiv preprint arXiv:2505.11581*, 2025.

Z. Kurth-Nelson, T. Behrens, G. Wayne, K. Miller, L. Luettgau, R. Dolan, Y. Liu, and P. Schwartenbeck. Replay and compositional computation. *Neuron*, 111(4):454–469, 2023.

Z. Kurth-Nelson, S. Sullivan, J. Z. Leibo, and M. Guitart-Masip. Dynamic diversity is the answer to proxy failure. *Behavioral and Brain Sciences*, 47:e77, 2024.

K. K. Ladha. The condorcet jury theorem, free speech, and correlated votes. *American Journal of Political Science*, pages 617–634, 1992.

B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015. doi: 10.1126/science. aab3050.

G. Lakoff and M. Johnson. *Metaphors We Live By*. University of Chicago Press, 1980.

N. Lane. *Vital question: energy, evolution, and the origins of complex life*. WW Norton & Company, 2015.

N. Lane and W. Martin. The energetics of genome complexity. *Nature*, 467(7318):929–934, 2010.

C. G. Langton. Computation at the edge of chaos: Phase transitions and emergent computation. *Physica D: nonlinear phenomena*, 42(1-3):12–37, 1990.

B. Latour. *We have never been modern*. Harvard university press, 2012.

S. Lazar and A. Nelson. Ai safety on whose terms?, 2023.

C. R. Leana. A partial test of janis' groupthink model: Effects of group cohesiveness and leader behavior on defective decision making. *Journal of management*, 11(1):5–18, 1985.

J. Lehtonen, M. D. Jennions, and H. Kokko. The many costs of sex. *Trends in ecology & evolution*, 27(3):172–178, 2012.

H. Lerner. *The dance of anger*. Harper & Row, 1985.

J. K. Leutgeb, S. Leutgeb, M.-B. Moser, and E. I. Moser. Pattern separation in the dentate gyrus and ca3 of the hippocampus. *science*, 315(5814):961–966, 2007.

C. Lévi-Strauss. *Les structures élémentaires de la parenté*. Presses Universitaires de France, Paris, 1949.

S. Lewandowsky, U. K. Ecker, C. M. Seifert, N. Schwarz, and J. Cook. Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest*, 13(3):106–131, 2012.

J. E. Lisman and O. Jensen. The theta-gamma neural code. *Neuron*, 77(6):1002–1016, 2013.

N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

A. Livnat, C. Papadimitriou, J. Dushoff, and M. W. Feldman. A mixability theory for the role of sex in evolution. *Proceedings of the National Academy of Sciences*, 105(50):19803–19808, 2008.

A. Livnat, C. Papadimitriou, N. Pippenger, and M. W. Feldman. Sex, mixability, and modularity. *Proceedings of the National Academy of Sciences*, 107(4):1452–1457, 2010.

G. Loewenstein. Out of control: Visceral influences on behavior. *Organizational Behavior and Human Decision Processes*, 65(3):272–292, 1996.

L. Luettgau, V. Cheung, M. Dubois, K. Juechems, J. Bergs, H. Davidson, B. O'Dell, H. R. Kirk, M. Rollwage, and C. Summerfield. People readily follow personal advice from ai but it does not improve their well-being. *arXiv preprint arXiv:2511.15352*, 2025.

M. C. Mackey and L. Glass. Oscillation and chaos in physiological control systems. *Science*, 197(4300):287–289, 1977. doi: 10.1126/science.267326.

L. Margulis and D. Sagan. *Microcosmos: Four Billion Years of Microbial Evolution*. Summit Books, New York, 1986.

L. Margulis and D. Sagan. *What Is Life?* Simon and Schuster, New York, 1995.

D. Marr. Simple memory: a theory for archicortex. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 262(841):23–81, 1971. doi: 10.1098/rstb.1971.0078.

W. Martin, J. Baross, D. Kelley, and M. J. Russell. Hydrothermal vents and the origin of life. *Nature Reviews Microbiology*, 6(11):805–814, 2008.

A. H. Maslow. A theory of human motivation. *Psychological review*, 50(4):370, 1943.

W. A. Mason, A. Jones, and R. L. Goldstone. Propagation of innovations in networked groups. *Journal of Experimental Psychology: General*, 137(3):422, 2008.

J. Maynard Smith. The origin and maintenance of sex. In *Group selection*, pages 163–175. Aldine Atherton, 1971.

J. Maynard Smith. *The evolution of sex*, volume 4. Cambridge University Press Cambridge, 1978.

M. McCain, R. Linthicum, C. Lubinski, A. Tamkin, S. Huang, M. Stern, K. Handa, E. Durmus, T. Neylon, S. Ritchie, K. Jagadish, P. Maheshwary, S. Heck, A. Sanderford, and D. Ganguli. How people use Claude for support, advice, and companionship. Anthropic, June 2025. URL https://www.anthropic.com/news/how-people-use-claude-for-support-advice-and-companionship.

J. L. McClelland, B. L. McNaughton, and R. C. O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419, 1995.

B. L. McNaughton and R. G. M. Morris. Hippocampal synaptic enhancement and information storage within a distributed memory system. *Trends in Neurosciences*, 10(10):408–415, 1987. doi: 10.1016/0166--2236(87)90011--8.

M. Merleau-Ponty. *Phénoménologie de la perception*. Gallimard, Paris, 1945.

L. Messeri and M. J. Crockett. Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627(8002):49–58, 2024.

E. K. Miller and J. D. Cohen. An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, 24(1):167–202, 2001.

G. A. Miller, G. Eugene, and K. H. Pribram. *Plans and the Structure of Behaviour*. Routledge, 1960.

S. Milli, D. Hadfield-Menell, A. Dragan, and S. Russell. Should robots be obedient? *arXiv preprint arXiv:1705.09990*, 2017.

M. Minsky. *Society of Mind*. Simon and Schuster, New York, NY, 1986.

S. Minuchin. *Families and Family Therapy*. Harvard University Press, 1974.

A. L. Mishara. Klaus conrad (1905–1961): Delusional mood, psychosis, and beginning schizophrenia. *Schizophrenia Bulletin*, 36(1):9–13, 2010.

A. Miyake, N. P. Friedman, M. J. Emerson, A. H. Witzki, A. Howerter, and T. D. Wager. The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive psychology*, 41(1):49–100, 2000.

H. Morrin, L. Nicholls, M. Levin, J. Yiend, U. Iyengar, F. DelGuidice, S. Bhattacharyya, J. MacCabe, S. Tognin, and R. Twumasi. Delusions by design? how everyday ais might be fuelling psychosis (and what can be done about it). 2025.

H. J. Muller. Some genetic aspects of sex. *The American Naturalist*, 66(703):118–138, 1932.

P. M. Müller, G. Miron, M. Holtkamp, and C. Meisel. Critical dynamics predicts cognitive performance and provides a common framework for heterogeneous mechanisms impacting cognition. *Proceedings of the National Academy of Sciences*, 122(14):e2417117122, 2025.

R. U. Muller and J. L. Kubie. The effects of changes in the environment on the spatial firing of hippocampal complex-spike cells. *The Journal of Neuroscience*, 7(7):1951–1968, 1987. doi: 10.1523/JNEUROSCI.07--07-01951.1987.

I. Murdoch and M. Midgley. *The sovereignty of good*. Routledge, 2013.

L. Myllyaho, M. Raatikainen, T. Männistö, T. Mikkonen, and J. K. Nurminen. Systematic literature review of validation methods for ai systems. *arXiv preprint arXiv:2107.12190*, 2021.

T. Nagel. The fragmentation of value. In *Mortal Questions*, pages 128–141. Cambridge University Press, Cambridge, 1979.

T. Nagel. *The View from Nowhere*. Oxford University Press, New York, 1986. ISBN 9780195056440.

F. Nietzsche. *Also sprach Zarathustra: Ein Buch für Alle und Keinen.* Ernst Schmeitzner, Chemnitz, 1883. Published in four parts. Parts 1–3 (1883–1884) by Schmeitzner; Part 4 (1885) privately printed by the author.

E. Nisbet and N. Sleep. The habitat and nature of early life. *Nature*, 409(6823):1083–1091, 2001.

M. A. Nowak and H. Ohtsuki. Prevolutionary dynamics and the origin of evolution. *Proceedings of the National Academy of Sciences*, 105(39):14924–14927, 2008.

S. Noy and W. Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192, 2023.

M. C. Nussbaum. *The fragility of goodness: Luck and ethics in Greek tragedy and philosophy.* Cambridge University Press, 2001.

J. P. O'Doherty, J. Cockburn, and W. M. Pauli. Learning, reward, and decision making. *Annual review of psychology*, 68(1):73–100, 2017.

S. Ohlsson. Information-processing explanations of insight and related phenomena. *Advances in the psychology of thinking*, 1:1–44, 1992.

C. O'Keefe, P. Cihon, B. Garfinkel, C. Flynn, J. Leung, and A. Dafoe. The windfall clause: Distributing the benefits of ai for the common good. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 327–331, 2020.

C. Olah, N. Cammarata, L. Schubert, G. Goh, M. Petrov, and S. Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

OpenAI. Democratic inputs to ai. `https://openai.com/blog/democratic-inputs-to-ai/`, May 2023. Accessed: 2025-12-05.

OpenAI. The state of enterprise AI 2025 report. Technical report, OpenAI, 2025. URL `https://cdn.openai.com/pdf/7ef17d82-96bf-4dd1-9df2-228f7f377a29/the-state-of-enterprise-ai_2025-report.pdf`.

L. D. Ordóñez, M. E. Schweitzer, A. D. Galinsky, and M. H. Bazerman. Goals gone wild: The systematic side effects of overprescribing goal setting. *Academy of Management Perspectives*, 23(1):6–16, 2009.

R. C. O'Reilly, T. E. Hazy, J. Mollick, P. Mackie, and S. Herd. Goal-driven cognition in the brain: a computational framework. *arXiv preprint arXiv:1404.7591*, 2014.

S. D. Østergaard. Generative artificial intelligence chatbots and delusions: From guesswork to emerging cases, 2025.

M. Owen. How to avoid the problem of 'group-think' in your boardroom, December 2019. URL `https://owenmorrispartnership.com/how-to-avoid-the-problem-of-group-think-in-your-boardroom/`.

V. Padmakumar and H. He. Does writing with language models reduce content diversity? *arXiv preprint arXiv:2309.05196*, 2023.

E. Perez, S. Ringer, K. Lukošiūtė, K. Nguyen, E. Chen, S. Heiner, C. Pettit, C. Olsson, S. Kundu, S. Caduri, et al. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022.

F. Perls, R. Hefferline, and P. Goodman. *Gestalt Therapy: Excitement and Growth in the Human Personality*. Julian Press, 1951.

D. Pimentel, R. Zuniga, and D. Morrison. Update on the environmental and economic costs associated with alien-invasive species in the united states. *Ecological economics*, 52(3):273–288, 2005.

S. Pinker. *The Language Instinct: How the Mind Creates Language*. William Morrow and Company, New York, 1994.

K. Polanyi. *The great transformation: The political and economic origins of our time*. Beacon Press, 1944.

M. Polanyi. *The Tacit Dimension*. Doubleday, Garden City, NY, 1966.

E. Polster and M. Polster. *Gestalt therapy integrated: Contours of theory & practice*, volume 6. Vintage, 1974.

K. R. Popper. *Logik der Forschung: Zur Erkenntnistheorie der modernen Naturwissenschaft*. Verlag von Julius Springer, Wien (Vienna), 1934.

Y. Potter, S. Lai, J. Kim, J. Evans, and D. Song. Hidden persuaders: Llms' political leaning and their influence on voters. *arXiv preprint arXiv:2410.24190*, 2024.

I. Prigogine and I. Stengers. *Order Out of Chaos: Man's New Dialogue with Nature*. Bantam Books, New York, 1984.

M. I. Rabinovich, R. Huerta, P. Varona, and V. S. Afraimovich. Transient cognitive dynamics, metastability, and decision making. *PLoS Computational Biology*, 4(5):e1000072, 2008. doi: 10.1371/journal.pcbi.1000072.

P. Railton. Moral realism. *The Philosophical Review*, 95(2):163–207, 1986.

D. M. Raup. The role of extinction in evolution. *Proceedings of the National Academy of Sciences*, 91(15):6758–6763, 1994.

J. Rawls. *A Theory of Justice*. Harvard University Press, Cambridge, MA, 1st edition, 1971.

J. Rawls. *Political Liberalism*. Columbia University Press, New York, 1993. ISBN 9780231052481.

W. Reich. On character analysis. *The Psychoanalytic Review (1913-1957)*, 20:89, 1933.

L. Reynolds and K. McDonell. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems*, pages 1–7, 2021.

W. J. Ripple and R. L. Beschta. Trophic cascades in yellowstone: the first 15 years after wolf reintroduction. *Biological Conservation*, 145(1):205–213, 2012.

J. Rockström, W. Steffen, K. Noone, Å. Persson, F. S. Chapin III, E. F. Lambin, T. M. Lenton, M. Scheffer, C. Folke, H. J. Schellnhuber, et al. A safe operating space for humanity. *nature*, 461(7263):472–475, 2009.

F. Roux and P. J. Uhlhaas. Working memory and neural oscillations: alpha–gamma versus theta–gamma codes for distinct wm information? *Trends in cognitive sciences*, 18(1):16–25, 2014.

S. Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin Publishing Group, 2019. ISBN 9780525558620. URL https://books.google.co.uk/books?id=M1eFDwAAQBAJ.

P. Saffo. Strong opinions weakly held, 2008. URL https://saffo.com/02008/07/26/strong-opinions-weakly-held/.

S. Santurkar, E. Durmus, F. Ladhak, C. Lee, P. Liang, and T. Hashimoto. Whose opinions do language models reflect? In *International conference on machine learning*, pages 29971–30004. PMLR, 2023.

C. B. Saper and B. B. Lowell. The hypothalamus. *Current Biology*, 24(23):R1111–R1116, 2014.

J. Sapp. *Evolution by association: a history of symbiosis*. New York: Oxford University Press, 1994.

B. T. Saunders and T. E. Robinson. The role of dopamine in the accumbens core in the expression of pavlovian-conditioned responses. *European Journal of Neuroscience*, 36(4):2521–2532, 2012.

D. L. Schacter, D. R. Addis, and R. L. Buckner. Remembering the past to imagine the future: the prospective brain. *Nature reviews neuroscience*, 8(9):657–661, 2007.

R. C. Schank and R. P. Abelson. *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology Press, 1977.

P. Scharre. *Army of None: Autonomous Weapons and the Future of War*. W. W. Norton, New York, 2018.

D. Schluter. *The ecology of adaptive radiation*. OUP Oxford, 2000.

E. Schrödinger. *What is Life? The Physical Aspect of the Living Cell*. Cambridge University Press, Cambridge, UK, 1944. Based on lectures delivered at Trinity College, Dublin, February 1943.

J. Schulkin and P. Sterling. Allostasis: a brain-centered, predictive mode of physiological regulation. *Trends in neurosciences*, 42(10):740–752, 2019.

J. C. Scott. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press, New Haven, CT, 1998. ISBN 9780300070163.

O. Seehausen, R. K. Butlin, I. Keller, C. E. Wagner, J. W. Boughman, P. A. Hohenlohe, C. L. Peichel, G.-P. Saetre, C. Bank, Å. Brännström, et al. Genomics and the origin of species. *Nature Reviews Genetics*, 15(3):176–192, 2014.

A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 59–68, 2019.

A. N. Sell. The recalibrational theory and violent anger. *Aggression and violent behavior*, 16 (5):381–389, 2011.

N. Selwyn. On the limits of artificial intelligence (ai) in education. *Nordisk tidsskrift for pedagogikk og kritikk*, 10(1):3–14, 2024.

T. V. Sewards and M. A. Sewards. Representations of motivational drives in mesial cortex, medial thalamus, hypothalamus and midbrain. *Brain research bulletin*, 61(1):25–49, 2003.

J. Y. Shah, R. Friedman, and A. W. Kruglanski. Forgetting all else: on the antecedents and consequences of goal shielding. *Journal of personality and social psychology*, 83(6):1261, 2002.

R. Shah, P. Freire, N. Alex, R. Freedman, D. Krasheninnikov, L. Chan, M. D. Dennis, P. Abbeel, A. Dragan, and S. Russell. Benefits of assistance over reward learning. *NeurIPS*, 2020.

M. Shanahan, K. McDonell, and L. Reynolds. Role play with large language models. *Nature*, 623(7987):493–498, 2023.

D. Shapiro. *Neurotic styles*. Basic Books, 1965.

M. Sharma, M. Tong, T. Korbak, D. Rogers, N. Bennett, Gabriel andjz, and E. Perez. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

M. Sharp, O. Bilgin, I. Gabriel, and L. Hammond. Agentic inequality. *arXiv preprint arXiv:2510.16853*, 2025.

W. L. Shew, H. Yang, T. Petermann, R. Roy, and D. Plenz. Neuronal avalanches imply maximum dynamic range in cortical networks at criticality. *Journal of neuroscience*, 29(49): 15595–15600, 2009.

W. L. Shew, H. Yang, S. Yu, R. Roy, and D. Plenz. Information capacity and transmission are maximized in balanced cortical networks with neuronal avalanches. *Journal of Neuroscience*, 31(1):55–63, 2011. doi: 10.1523/JNEUROSCI.4637--10.2011.

V. Shiva. *Monocultures of the mind: Perspectives on biodiversity and biotechnology*. Palgrave Macmillan, 1993.

J. Sidanius and F. Pratto. *Social dominance: An intergroup theory of social hierarchy and oppression*. Cambridge University Press, 2001.

A. M. Sillito. The contribution of inhibitory mechanisms to the receptive field properties of neurones in the striate cortex of the cat. *The Journal of Physiology*, 250(2):305–329, 1975.

P. Singer. *The expanding circle*. Clarendon Press Oxford, 1981.

A. Singla, A. Sukharevsky, L. Yee, M. Chui, B. Hall, and T. Balakrishnan. The state of AI in 2025: Agents, innovation, and transformation. Technical report, McKinsey & Company, Nov. 2025. URL https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai.

N. H. Sleep. The hadean-archaean environment. *Cold spring harbor perspectives in biology*, 2 (6):a002527, 2010.

M. Sloane, E. Moss, O. Awomolo, and L. Forlano. Participation is not a design fix for machine learning. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–6, 2022.

A. Smith. *An inquiry into the nature and causes of the wealth of nations: Volume One*. London: printed for W. Strahan; and T. Cadell, 1776., 1776.

E. Smith and H. Morowitz. *The Origin and Nature of Life on Earth: The Emergence of the Fourth Geosphere*. Cambridge University Press, 2016.

M. Smith. *The Moral Problem*. Blackwell, Oxford, 1994.

M. J. Smith. *When I say no, I feel guilty*. Bantam, 1985.

P. Smolensky. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46(1–2):159–216, 1990. doi: 10.1016/0004--3702(90)90007-M.

N. Soares and B. Fallenstein. Aligning superintelligence with human interests: A technical research agenda. *Machine Intelligence Research Institute (MIRI) technical report*, 8, 2014.

S. Sontag, C. Drew, and A. L. Drew. *Blind man's bluff: The untold story of American submarine espionage*. Public Affairs, 1998.

T. Sorensen, J. Moore, J. Fisher, M. Gordon, N. Mireshghallah, C. M. Rytting, A. Ye, L. Jiang, X. Lu, N. Dziri, et al. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*, 2024.

D. Speijer, J. Lukeš, and M. Eliáš. Sex is a ubiquitous, ancient, and inherent attribute of eukaryotic life. *Proceedings of the National Academy of Sciences*, 112(29):8827–8834, 2015.

D. Sperber, F. Clément, C. Heintz, O. Mascaro, H. Mercier, G. Origgi, and D. Wilson. Epistemic vigilance. *Mind & language*, 25(4):359–393, 2010.

K. O. Stanley and J. Lehman. *Why Greatness Cannot Be Planned: The Myth of the Objective*. Springer, Cham, 2015. ISBN 978-3-319-15523-4.

K. O. Stanley, J. Lehman, and L. B. Soros. Open-endedness: The last grand challenge you've never heard of. O'Reilly Radar, 2017. URL https://www.oreilly.com/radar/open-endedness-the-last-grand-challenge-youve-never-heard-of/. Accessed: 2025-01-28.

G. Stasser and W. Titus. Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of personality and social psychology*, 48(6):1467, 1985.

R. Stein. *Incest and human love: The betrayal of the soul in psychotherapy*. Third Press, 1973.

R. J. Stern and T. V. Gerya. The importance of continents, oceans and plate tectonics for the evolution of complex life: implications for finding extraterrestrial civilizations. *Scientific Reports*, 14(1):8552, 2024.

M. Stocker. *Plural and conflicting values*. Clarendon Press, 1992.

S. G. Straus, A. M. Parker, and J. B. Bruce. The group matters: A review of processes and outcomes in intelligence analysis. *Group Dynamics: Theory, Research, and Practice*, 15(2): 128, 2011.

J. Surowiecki. *The wisdom of crowds*. Vintage, 2005.

D. Susskind. *A World Without Work: Technology, Automation, and How We Should Respond*. Metropolitan Books, New York, 2020.

R. I. Sutton and M. R. Louis. How selecting and socializing newcomers influences insiders. *Human Resource Management*, 26(3):347–361, 1987.

S. Suzuki. *Zen Mind, Beginner's Mind*. Weatherhill, New York, NY, 1970. Reference for 'Don't serve them tea'.

V. Tausk. Über die entstehung des 'beeinflussungsapparates' in der schizophrenie. *Internationale Zeitschrift für Psychoanalyse*, 5:1–33, 1919.

TechCrunch. Sam altman says ChatGPT has hit 800m weekly active users, Oct. 2025. URL https://techcrunch.com/2025/10/06/sam-altman-says-chatgpt-has-hit-800m-weekly-active-users/. Accessed: 2025-11-19.

B. J. Tepper. Consequences of abusive supervision. *Academy of management journal*, 43(2): 178–190, 2000.

P. E. Tetlock. A value pluralism model of ideological reasoning. *Journal of personality and social psychology*, 50(4):819, 1986.

N. Tiku and S. Malhi. What is 'ai psychosis' and how can chatgpt affect your mental health? *The Washington Post*, August 2025. URL https://www.washingtonpost.com/health/2025/08/19/ai-psychosis-chatgpt-explained-mental-health/.

D. Tilman. *Resource competition and community structure*. Number 17. Princeton university press, 1982.

D. Tilman. Biodiversity: population versus ecosystem stability. *Ecology*, 77(2):350–363, 1996.

E. Tognoli and J. S. Kelso. The metastable brain. *Neuron*, 81(1):35–48, 2014.

J. Tooby and L. Cosmides. The psychological foundations of culture. In J. H. Barkow, L. Cosmides, and J. Tooby, editors, *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, pages 19–136. Oxford University Press, New York, NY, 1992.

D. S. Touretzky and G. E. Hinton. A distributed connectionist production system. *Cognitive Science*, 12(3):423–466, 1988. doi: 10.1207/s15516709cog1203_3.

H. M. Trainer, J. M. Jones, J. G. Pendergraft, C. K. Maupin, and D. R. Carter. Team membership change "events": A review and reconceptualization. *Group & Organization Management*, 45(2):219–251, 2020.

A. Treves and E. T. Rolls. Computational analysis of the role of the hippocampus in memory. *Hippocampus*, 4(3):374–391, 1994. doi: 10.1002/hipo.450040319.

A. M. Turing. The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 237(641):37–72, 1952. doi: 10.1098/rstb. 1952.0012.

A. M. Turner, D. Hadfield-Menell, and P. Tadepalli. Conservative agency via attainable utility preservation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 385–391, 2020.

N. Vafeas. Length of board tenure and outside director independence. *Journal of Business Finance & Accounting*, 30(7-8):1043–1064, 2003.

R. R. Vallacher and D. M. Wegner. What do people think they're doing? action identification and human behavior. *Psychological review*, 94(1):3, 1987.

M. Van Der Meer, Z. Kurth-Nelson, and A. D. Redish. Information processing in decision-making systems. *The Neuroscientist*, 18(4):342–359, 2012.

F. J. Varela, E. Thompson, and E. Rosch. *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press, Cambridge, MA, 1991.

S. L. Videbeck. *Psychiatric-mental health nursing*. Lippincott Williams & Wilkins, 2010.

N. Virgo, C. Fernando, B. Bigge, and P. Husbands. The elongation catastrophe in physical self-replicators (full article). In *ECAL 2011: The 11th European Conference on Artificial Life*. Citeseer, 2011.

L. Von Bertalanffy. The theory of open systems in physics and biology. *Science*, 111(2872): 23–29, 1950.

G. Wächtershäuser. Before enzymes and templates: theory of surface metabolism. *Microbiological reviews*, 52(4):452–484, 1988.

G. P. Wagner and L. Altenberg. Perspective: complex adaptations and the evolution of evolvability. *Evolution*, 50(3):967–976, 1996.

B. Walker, C. S. Holling, S. R. Carpenter, and A. Kinzig. Resilience, adaptability and transformability in social–ecological systems. *Ecology and society*, 9(2), 2004.

H. Watson. Biological membranes. *Essays in biochemistry*, 59:43–69, 2015.

A. Watts. *The Wisdom of Insecurity: A Message for an Age of Anxiety*. Vintage Books, New York, NY, 2011. Original work published 1951; Reference for The Backwards Law.

D. M. Wegner. Ironic processes of mental control. *Psychological review*, 101(1):34, 1994.

V. Weilnhammer, K. Y. Hou, R. Dolan, and M. M. Nour. Vulnerability-amplifying interaction loops: a systematic failure mode in ai chatbot mental-health interactions. *arXiv preprint arXiv:2602.01347*, 2026.

A. Weismann. *Essays upon heredity and kindred biological problems*. Clarendon Press, Oxford, 1889.

P. Welinder, S. Branson, P. Perona, and S. Belongie. The multidimensional wisdom of crowds. *Advances in neural information processing systems*, 23, 2010.

N. Wiener. Some moral and technical consequences of automation: As machines learn they may develop unforeseen strategies at rates that baffle their programmers. *Science*, 131(3410): 1355–1358, 1960.

K. Wilber. *The Atman Project: A Transpersonal View of Human Development.* Quest Books, Wheaton, IL, 1996. Reference for The Subject-Object Shift.

B. Williams. Conflicts of values. In *Moral Luck: Philosophical Papers 1973–1980*, pages 71–82. Cambridge University Press, Cambridge, 1981.

B. Williams. *Ethics and the Limits of Philosophy.* Harvard University Press, Cambridge, MA, 1985.

G. C. Williams. *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought.* Princeton University Press, Princeton, NJ, 1966.

L. Wittgenstein. *Tractatus Logico-Philosophicus.* Routledge & Kegan Paul, London, 1922.

M. J. Wood, K. M. Douglas, and R. M. Sutton. Dead and alive: Beliefs in contradictory conspiracy theories. *Social psychological and personality science*, 3(6):767–773, 2012.

S. C. Woolley and P. N. Howard. *Computational propaganda: Political parties, politicians, and political manipulation on social media.* Oxford University Press, 2018.

S. Wu, B. A. Nijstad, and Y. Yuan. Membership change, idea generation, and group creativity: A motivated information processing perspective. *Group Processes & Intergroup Relations*, 25 (5):1412–1434, 2022.

T. Wu. *The master switch: The rise and fall of information empires.* Vintage, 2011.

R. Xu, Z. Qi, Z. Guo, C. Wang, H. Wang, Y. Zhang, and W. Xu. Knowledge conflicts for llms: A survey. *arXiv preprint arXiv:2403.08319*, 2024.

W. Xu, N. Jojic, S. Rao, C. Brockett, and B. Dolan. Echoes in ai: Quantifying lack of plot diversity in llm outputs. *Proceedings of the National Academy of Sciences*, 122(35):e2504966122, 2025.

S. Yachi and M. Loreau. Biodiversity and ecosystem productivity in a fluctuating environment: the insurance hypothesis. *Proceedings of the National Academy of Sciences*, 96(4):1463–1468, 1999.

R. V. Yampolskiy. On monitorability of ai. *AI and Ethics*, 5(1):689–707, 2025.

J. A. Yeung, J. Dalmasso, L. Foschini, R. J. Dobson, and Z. Kraljevic. The psychogenic machine: Simulating ai psychosis, delusion reinforcement and harm enablement in large language models. *arXiv preprint arXiv:2509.10970*, 2025.

G. M. Yontef. *Awareness, dialogue & process: Essays on Gestalt therapy.* The Gestalt Journal Press, 1993.

N. Yousif. Parents of teenager who took his own life sue openai, August 27 2025. URL `https://www.bbc.com/news/articles/cgerwp7rdlvo`.

E. Yudkowsky. Coherent extrapolated volition. *Singularity Institute for Artificial Intelligence*, 2004.

Z. Zhao, E. Wallace, S. Feng, D. Klein, and S. Singh. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR, 2021.

E. Zhou and D. Lee. Generative artificial intelligence, human creativity, and art. *PNAS nexus*, 3(3):pgae052, 2024.

S. Zhuang and D. Hadfield-Menell. Consequences of misaligned ai. *Advances in Neural Information Processing Systems*, 33:15763–15773, 2020.