Alignment, Boundaries, Contextualization

November 21, 2025

In this paper, we begin by surveying a range of natural and human systems and observing commonalities. Each form – an equation, a genome, a viewpoint, an institution – is myopic in the sense that it is only a small aspect of the world. But the existence of many partial forms, developing rich, grounded individual identities while flexibly working together at many scales, constitutes the livingness of the world. Collapse is a failure mode where, instead of lightly-held dynamic relationships between different partialities, there is overcommitment to particular forms. Cancer spreads, concentrated power reduces human welfare, invasive species choke out complex ecosystems. In living systems, semi-permeable boundaries protect against collapse by maintaining distinctiveness of separate forms while also allowing them to interact productively. Boundaries thereby contextualize forms in relation to others, surfacing paradoxes by juxtaposing incompatible things. Boundaries underpin nuance and the continued arising of unexpected phenomena.

We then use those living systems as inspiration for a new way to think about AI alignment. We propose that alignment is the ongoing process of limiting overcommitment to any form. On the path it is currently following, AI will charge particular myopic forms with enormous leverage, creating a unique risk of overcommitment. If aspects of the AI system remain fixed while it gains increasing resource, capability and purview, there is a possibility of severe collapse. Some researchers believe AI is inevitably destructive for these reasons. We cautiously argue that there is a non-destructive path available, by reframing AI as a continuation of living processes that bound one another. In this view, alignment is an evolving network of semi-permeable boundaries that contextualize any particular form of AI to avoid collapse and support a deepening of the mystery of life.

Contents

1 Examples from the natural and human world

The central ideas of this paper are necessarily abstract: we hope they may contribute to reasoning about AI even when it is vastly different from anything we are familiar with. To connect with these abstract ideas, we walk through a series of examples from living systems. Each example illustrates the core principle of the paper. In some examples we also drill deeper into subthemes that are especially vivid in that setting. Within each example, we hope the ideas are approachable if not commonsense. Seeing how the same patterns play out across a wide range of systems may also foreground their generality.

1.1 Cell membranes

'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

The cell membrane is an essential boundary of living systems. The membrane holds the integrity of the cell against the overwhelming pressure of diffusion that tries to homogenize the cell with the outside (Alberts et al., 2022; Bray, 2019; Harold, 2001; Lane, 2015; Watson, 2015).

The membrane places limits on interactions between the inside and the outside. Thanks to the membrane, both the cell and the outside can exist; this is a more diverse, less symmetric arrangement compared to the inside and outside being blended together (Anderson, 1972; Prigogine and Stengers, 1984; Schrödinger, 1944; Turing, 1952). Without boundaries, interactions cause collapse, where there are no longer separate entities flexibly interacting, but instead overcommitment to a simpler homogeneous form[1].

Cell membranes are semi-permeable: they prevent the conditions outside (neighboring cells or the extracellular space) 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. 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: this is 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 put to work the potential energy of the asymmetry between different forms. Without the membrane, the pressure of chemical gradients would rapidly homogenize the cell with the outside. With the membrane, the same gradients 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. This pattern is common across many kinds of systems and will be important for the alignment problem. We will return to it a few times.

Another recurring thread is that collapse is always relative. For example, programmed cell death is catastrophic collapse at the level of the dying cell, but it can be beneficial or even necessary for the organism the cell belongs to.

1.2 Group problem solving

"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

[1]We define collapse as overcommitment to particular forms. It could be equivalently defined as either undercommitment or overcommitment. Radically undercommitting means homogeneity, which is itself a particular kind of form and so also overcommitted.

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, the wisdom of crowds is diminished if the group lacks diversity, either ab initio or as a result of within-group communication and influence (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 often 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. There is therefore a tension between overcommunication where diversity is lost and undercommunication where diversity is not leveraged.

The crux is semi-permeable boundaries: wisely transmitting the right information at the right time, in the right way. Thoughtful strategies for communication are like the transmembrane channels that allow the right molecules in and out of the cell at the right time. They protect the existence of diverse problem solving approaches while also allowing productive interaction between them.

Many varieties of semi-permeable boundary are 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). In a later section, we will look at boundaries within an individual, such as skepticism, that make it easier to interact with others without overwriting one’s own beliefs.

A particularly important boundary for group problem solving is simply giving members the space to work independently before communicating (Frey and Van de Rijt, 2021; Surowiecki, 2005). In the case of the submarine search, experts weren’t allowed to communicate while forming their own estimates; the estimates were later aggregated in a principled way by Craven. Analogously, science historians argue that partial intellectual isolation has at times been beneficial for the emergence of deeply new ideas. Einstein’s relative independence from the advanced mathematical techniques of contemporaries like Hilbert led to a theory of general relativity grounded in deep physical insight rather than mathematical convenience (Corry et al., 1997; Renn and Sauer, 1999; Stachel, 1989). Newton's and Leibniz's famous independent development of calculus, as a result of their mutual isolation, yielded two distinct and valuable mathematical systems that complemented and enriched one another (Hall, 2002).

The benefit of temporary isolation before communicating also shows up in controlled experiments. Bernstein et al. (2018) tasked small groups with solving instances of the traveling salesman problem. Each group was randomly assigned to one of three conditions. In some groups, members could continually see the work of other members as they progressed toward a solution; in some groups members could only occasionally exchange progress; and in some groups there was no exchange. The researchers found that groups with continual information exchange rarely found good solutions. In these groups, typically one individual would stumble on a solution that looked compelling but was actually a dead-end. When this solution was immediately shared with others, it hampered their progress. Groups with occasional or no contact were much more likely to find optimal or near-optimal solutions.

We stress that this is not an indictment of connection and communication between group members. Rapid access to information and shared solutions often demonstrably boosts productivity. In some situations the ideal boundary might be working in isolation for months at a time. But in other situations it could be daily meetings with intensive communication, while maintaining the self-confidence to keep pursuing one's own intuition in the face of skepticism from others (Paulus and Nijstad, 2003; Sawyer, 2017). The key is that boundaries support flexible interactions and avoid overcommitment to particular forms.

Structured space	**Force**	**Outcome without boundary**	**Semi-permeable boundary**	**Outcome if potential is held by boundary**
Competing drives and goals in an organism	Drive to eat	Obesity	Other drives, self-control, supportive environmental systems	Nutritional needs satisfied without overeating
Complex ecosystem	Human drive for expansion	Resource depletion, mass extinction	Measured regulatory policy	Economic growth without extensive ecosystem destruction
Individuals have different identities and motives	P's will to dominate	Loss of agency in Q	Owned anger in Q	Relating while maintaining individual autonomy
An intricate, balanced economy	Profit motive of one company	Monopoly and reduced innovation	Laws that allow profit seeking within limits	Productive competition
Multiple perspectives within an individual	Diffusion and drive for simplicity	Collapse to rigid thinking	Recognition of uncertainty	Beliefs that are stable but also adaptive and evolving
Distinct intra- and extra-cellular environments	Elecrochemical gradients	Dissolution of cell	Cell membrane	Cell maintains integrity but also processes external signals
Orderly cell types and tissues	Mutation and selection on cell lineages	Cancer	DNA repair, tumor suppression	Cancer is minimized while mutations can still benefit immunity and germ-line evolution
Individuals have different problem-solving methods	Social conformity, diffusion of ideas	Groupthink	Thinking separately before sharing results	Wisdom of crowds
Rich array of representations in the brain	Diffusion to equilibrium	Blending of representations	Lateral inhibition	Separate representations exist but can also interact

Table 1: Mapping some example systems into our terminology.

1.3 Genes

'The mere act of crossing by itself does no good. The good depends on the individuals which are crossed differing slightly in constitution, owing to their progenitors having been subjected during several generations to slightly different conditions.'
Charles Darwin

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; Speijer et al., 2015). Yet all known species either reproduce sexually or have some form of horizontal gene transfer (Butterfield, 2000; Gladyshev et al., 2008). Why is that?

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.

Sexual reproduction is a boundary that softens these 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". 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.

Boundaries encourage lightly-held, modular interactions. By not overcommitting to a particular genome, sex encourages genes to flexibly interact with other genes (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 deep 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.

1.4 Laws

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

'Growth for the sake of growth is the ideology of the cancer cell.'
Edward Abbey

Individual actors in a society and in an economy each act from their own perspective. Each actor's perspective is myopic because they cannot know everything or fully understand the motives and beliefs of others. Of course, myopia does not always mean selfishness in the sense of valuing only one's own wealth or physical wellbeing (Becker, 1974; Crockett et al., 2014; Henrich et al., 2001).

Without boundaries, one actor's perspective can dominate, resulting in collapse and 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 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.

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).

Of course, 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). Again, these evolving laws gradually acquire grounded wisdom as they are tested against many different situations and motives.

1.5 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 (Wikipedia, 2025).

Kirk stepped outside the mental frame in which there was an apparently unwinnable dilemma. From inside a particular frame, the frame appears to be reality. But there are almost always multiple valid perspectives, each of which is only a partial description of reality (Aristotle, 2019; De Bono, 1970; Duncker, 1945; Goffman, 1974; Heidegger, 1998; Javed and Sutton, 2024; Kant, 1781; Korzybski, 1933; Kuhn, 1970; Lakoff and Johnson, 1980; Ohlsson, 1992; Plato, 2002; Popper, 1934; Saffo, 2008; Wittgenstein, 1922). Famously, 'all models are wrong' (Box, 1976). Humans have a vast array of available metaphors and concepts, which are not even all consistent or compatible with one another (Adorno et al., 1950; Feyerabend, 1975; Freud, 1936; Hofstadter, 2001; Wood et al., 2012). The world is too complex for all beliefs to be fully evaluated against each other and reconciled. At any given time, we only access a very few items, and others are largely inaccessible (Baddeley, 2000; Dehaene, 2014; Hills et al., 2015; Miller, 1956). Each particular frame or concept is myopic because it doesn't capture the whole world, but collectively they form a powerful toolkit for problem solving and understanding.

The capacity to adopt multiple perspectives is, fittingly, described in multiple ways across different areas of psychology and cognitive science. 'Psychological flexibility' is the ability to update one's approach or lens contextually rather than being fused to a single thought or frame (Cherry et al., 2021). Conversely, 'functional fixedness' is excess attachment to one perspective (Duncker and Lees, 1945). 'Adaptive experts' dynamically evaluate the appropriateness of different interpretations, analogies or schemas (Feltovich et al., 1997; Hatano and Inagaki, 1984; Spiro et al., 1988). 'Integrative complexity' is first differentiating multiple perspectives on a problem and then identifying connections between them (Suedfeld et al., 1992; Tetlock, 1986). Humans contextually switch between many 'heuristics', each of which processes a problem through its own narrow lens (Gigerenzer and Brighton, 2009). 'Set shifting' is transitioning between task sets, which are the concepts and lenses relevant for particular tasks (Grant and Berg, 1948; Miyake et al., 2000). These psychological constructs capture a range of scales: people can hold multiple perspectives on something as fine-grained as the color of a dress or something as all-encompassing as their self-construct and the nature of reality.

Losing the ability to flexibly shift between different frames or thought patterns runs the risk of obsession or delusion. In obsession, a particular thought pattern or schema is overemphasized to the detriment of healthy functioning (Rachman, 1998; Salkovskis, 1985). In delusions, an entire conceptual framework crystallizes with excessive certainty and is resistant to disconfirmatory evidence (Adams et al., 2013; APA, 2013; Heinz et al., 2019; Jaspers, 1997; Mishara, 2010). Obsessions and delusions are myopic: they lose sight of most of the world by overcommitting one thought pattern or frame.

We stay flexible using the internal boundary of holding our own ideas lightly. As a playful example, author Lisa Stardust claims that "the moon controls the tides of the ocean, and we are made of 60 percent water. This means that the moon has a huge effect on all of us" (Mitchell, 2021). You probably immediately spotted the flaw in this argument. But at a zeroth order level, the argument does make perfect sense: W impacts X, X is made of Y, Z is also made of Y, so W should impact Z. Overriding this logic requires a higher order correction term: tides arise from differential tugging over long distances in a body of water that is free to slosh around. Adding the correction term is an increase in subtlety. Subtle correction terms are often hard-won knowledge originating from thoughtful interactions with the world. But we only profit from those interactions if we accept that our current model isn't the final answer[2]. As our ideas are tested against multiple situations and problems, they are refined and take on some of the deep structure of the world, a grounded wisdom.

Crucially, the existence of narrow points of view is not a problem. It's necessary. All points of view are partial. Even obsession can be powerful when we obsess on a problem at work and occasionally achieve good results. A delusion-like framework can seed a scientific revolution. The point is not to shut down narrow concepts. The point is to limit them from becoming the sole and absolute determinants of behavior. I might work obsessively on a project while also having a rule that I must go to bed at 10 pm. This is a semi-permeable boundary. It doesn't block me from temporarily taking a strong perspective, but it does place contextual limits on it. When boundaries are semi-permeable, 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). Semi-permeable boundaries situate myopic frames within a larger context.

1.6 Ecosystems

An ecosystem's health and resilience depend on boundaries that limit the effectiveness of any constituent organism or group (Holling et al., 1973). Each entity tries to consume resources and proliferate, but if it succeeds too thoroughly, the ecosystem suffers.

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. Predation supplied a semi-permeable boundary: it placed contextualizing limits on the elk, without preventing them from fighting for their own survival and flourishing. The elk, by trying to optimize their own objectives within a broader context, 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).

Healthy ecosystems constitute a large evolving network of reciprocal or otherwise cyclical bound-

[2]Boundaries also protect Stardust's mystic beliefs. Boundaries create space for the mystic frame to explore its own reality. Stardust doesn't know a priori how right or wrong the mystic frame is; sometimes we need space to explore ideas everyone else thinks are crazy, like heliocentrism. Even *after* Stardust discovers that the mystic frame doesn't do well predicting a large class of sensory evidence, she can still hold it as a frame that has some value – perhaps it resonates with some internal psychological structure, like Jungian archetypes. If nothing else, remembering the internal logic of that frame might help her empathize with others who believe it. Contextualization holds the mystic frame for what it is, while simultaneously understanding that the Newtonian explanation is better for launching projectiles.

aries between the many players: predation, parasitism, resource competition and so on. Boundaries drive the evolution of new structure. For example, competition leads to niche partitioning, where species evolve to use different resources or the same resources in different ways, increasing ecosystem complexity and resilience (Schoener, 1974). The myopic motives of each species, when contextualized by semi-permeable boundaries, work toward open-ended enriching of life.

Human drives within ecosystems are 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 lack of boundaries can result in overcommitment to those aims, with a negative impact on both ecosystem health and our own welfare.

Fortunately, there are some boundaries on human actions within ecosystems. One is our own finite capability. Another is that excessively extractive civilizations sometimes fail and are replaced by longer-sighted ones (Diamond, 2004). In recent times, the effectiveness of these two boundaries has waned because our capabilities are increasing and we're becoming a single global civilization. But through the long-sightedness of intelligence, we sometimes foresee the consequences of excess extraction and place our own limits on it, including state regulation, self-policing and environmental certifications. These self-imposed boundaries are productive because they are semi-permeable. Regulation does not forbid the extraction of all resources. It places contextual limits in response to information about our resource needs as well as what is sustainable (Lazarus, 2023). Interestingly, our long-sighted intelligence arose from short-sighted evolution.

Finally, we again stress that one entity's collapse is 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 even when we try to be responsible; yet the extraction fuels explosion of technology, art, music, and human experience.

1.7 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.

1.8 Information in the brain

'When I observe something unusual in an experiment, it reverberates in my brain for a long while.'
György Buzsáki

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

The brain is somewhat miraculous in keeping so 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 this 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 that 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 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). The capacity to separate things into nearly-discrete entities and then recombine them in vast numbers of structured ways powers the extraordinary human capacity for reasoning (Chomsky, 1957; Fodor, 1975; Kurth-Nelson et al., 2023; Lake et al., 2015; Pinker, 1994). Again, semi-permable boundaries keep forms distinct while enabling them to flexibly and modularly interact.

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 Motivation

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 severe health risks. Single-minded pursuit of sex causes relational, occupational, legal and health harms (Carnes, 2001; Kraus et al., 2016). Overcommitment to a single drive means the organism becomes unwell.

The space 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 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 both time and space (Ballard et al., 2018; Evenden, 1999; Shah et al., 2002; Vallacher and Wegner, 1987). Narrow in time means being focused 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 is at the expense of a broader balance of goals – and at the expense of the health of the organism or other individuals. We suggest that health could reasonably be defined as not overcommitting to a particular form.

Overcommiting to a particular strategy for satisfying a drive or goal can even come at the expense of satisfying that very drive or goal. 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, food is there, so I'll go there", and thus failed to even satisfy the drive for food (Dayan et al., 2006; O'Doherty et al., 2017; Van Der Meer et al., 2012). The zeroth order logic recalls Lisa Stardust's model of physics from Section 1.5.

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. The chicken wants food and tries to take the shortest path toward it in the naive sense of a straight line through space. But 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.

Boundaries, on the other hand, translate the pressure of motivation into higher-order structure

– the best way to approach the food is not the shortest path in space. Instead, achieving the goal depends on discovering a new solution. Semi-permeable boundaries support formation of new structure by placing contextualizing limits.

A broad class of boundaries on particular drives, goals and strategies is *cognitive control* (Botvinick et al., 2001; Braver, 2012; Miller and Cohen, 2001; Miyake et al., 2000). In the case of overeating, control contextualizes the food-seeking drive. In the case of the chickens, control contextualizes the prepotent tendency to approach the food. In the case of over-focusing on a single goal like work, control helps with task switching. Cognitive control is a *semi-permeable* boundary: it does not erase particular goals, but instead contextualizes them within a larger system.

1.10 Contemplative practice

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

'Real love will take you far beyond yourself; and therefore real love will devastate you.'
Ken Wilber

Some forms exist in our minds without awareness. Think of an assumption somebody has that's never been questioned. That assumption could be life-long and self-defining, or it could be fleeting and perceptual, like the assumption that the thing I'm touching is a keyboard. Unquestioned assumptions are overcommitment. We believe in them inflexibly. But sometimes there's a moment of stepping back, where the assumed form becomes an object in awareness. In that moment, the assumption is contextualized. We realize it's not an absolute truth, but rather a form in our minds. Awareness is contextualization.

Contemplative traditions suggest that the only 'absolute' truth is the self-evident truth of immediate experience – awareness itself. Of course, even the concept of awareness is relative and infinitely incomplete. Once we picture awareness as an object, it's not the thing we're talking about. So the word 'truth' is not really describing any particular thing at all. We could use different language and describe it as something more like an orientation toward stepping back from each perspective into awareness. And again, any concept we have of that process is not what we're really talking about. By construction, contextualization is an unsolvable mystery from any particular point of view.

We could also think of awareness as an evolving system of boundaries. It's the process of limiting overcommitment to any thing. What it takes to limit overcommitment to A is different than what it takes for B, so new boundaries are needed as the situation changes. This will be relevant for AI alignment in the next section. The boundaries of awareness are semi-permeable because they don't reject the form they contextualize. 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. Awareness holds us at the knife's edge of not collapsing exclusively into any particular forms. This activates a deeper sensitivity to ourselves and to the world. Subtler forms, which would have been erased by overcommitment to other forms, instead play a role in a richer overall internal structure. Our own potential within the world creatively emerges in continued newness.

Contemplative philosophy posits that suffering comes from overcommitment to particular conceptualizations or desires: believing excessively in a formalism. Being attached to particular concepts, beliefs, feelings and other patterns in a collapsed way. There's always something we believe, something we can't even see as an object because it's so tautological for us. We keep trying to give ourselves what we think we want under this model, pretending that things are formalizable, but as a result we become less sensitive to the rest of the world. The parts of the world not covered by our concepts subjectively appear terrifying or morally wrong. And what we do to prevent the tautologically bad thing from happening is inevitably what causes the bad thing to continue. In other words, our collapsed patterns hold the tension that paradoxically creates the unease they resist[3].

But awareness contextualizes these dynamics. Stepping back into awareness can feel infinitely scary from the original frame, because it's potentially allowing the tautologically bad thing. But from the new frame, the bad thing is just another texture of experience, without being bad in an absolute sense. The fear or wrongness of not-self is no longer an absolute but instead exists in relationship. So awareness brings healing and growth. People often report subjectively that the energy locked in the darkness turned out to be full of life, and that there's something self-evidently good or beautiful about participating in this mysterious discovery of new structure and relationship.

Finally, we appreciate that this way of talking raises red flags for some people. In case a reframe might be helpful, the idea here isn't really any different from art. The orientation toward not collapsing into particular concepts is familiar in art, poetry, music, dance. The meaning of art is open-ended and changes with context – it has an inner life. What we value is perhaps something about 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

Now we take the lens we've been using to describe living systems and turn it to the AI alignment problem. Here's our central thesis: overcommitment to any particular form of AI means collapse, and AI carries a singular risk of overcommitment because of the extremities attached to it.

Let's break down what this means. First, what are some of the particular forms of AI? Here are some examples:

- A trained model with fixed weights.
- A set of training data.
- A reward function or goal.
- An RL algorithm.

[3]Some schools of thought go a step farther to observe that whatever our current self is, it is always already inevitably contextualized, and love has no opposite.

- A model architecture.
- An architectural template like the transformer.
- A learned representation within the model.
- An input/output format or modality.
- A set of benchmarks.
- A system prompt.
- Following pretraining with posttraining.
- A process for obtaining feedback from humans.
- A procedure for continual learning.
- Emergent reasoning patterns that are resistant to change even with continued training.
- An AI agent that polarizes against another AI agent.
- Global hardware deployments that build in assumptions about the kinds of operations that can be done efficiently.
- The wishes of an individual human filtered through a steerable AI system.
- Feedback loops with humans arising from anthropomorphization of AI.
- Systems of AI belief that perpetuate through inter-AI communication.

Of course, all technologies have their particular imperfect forms at any given moment in time. In 1900, boiler explosions killed an estimated two people per day in North America (CEP Forensic, 2021). Historically, humans have iterated on technology, doing our best today while knowing future generations will improve on our solutions.

But AI comes with a number of extremities with the potential to produce 'lock-in' or overcommitment. For example, if a model with a particular goal becomes superintelligent, it might successfully resist human efforts to later update that goal (Bostrom, 2014; Russell, 2019).

Some of the extremities attached to AI (with substantial overlap between categories):

- **Intelligence.** Our own intelligence has always given us the ability to control other technologies, and other creatures. For the first time, we're endowing something else with it.
- **Extraordinary capabilities.** In some domains AI already greatly exceeds human abilities, like in making sense of huge amounts of data, performing certain kinds of reasoning, storing encyclopedic knowledge.
- **Intensity of optimization.** Today's version of AI training consists of a massive gradient descent effort, bringing enormous resources to bear toward pushing trillions of numbers in a coherent direction.
- **Economic concentration.** AI spending will reach an estimated 1.7% of global GDP in 2026 (Gartner, 2025), rivaling the concentration of spending during the buildouts of

electrification, railroads, and automobiles (Fogel, 1964; Nye, 1992). We are building out physical infrastructure, using energy, deploying human intellect and metabolizing data at extraordinary scale.

- **Concentration of information flow and human attention.** As of October 2025, ChatGPT alone had 800 million weekly active users (TechCrunch, 2025), and global AI usage continues to grow rapidly. This means that a huge fraction of people around the world are ingesting the same source. It is an open question to what degree current AI chatbots reflect the whole diversity of the internet versus collapsing it to a much narrower manifold (Kirk et al., 2023; Xu et al., 2025). And of course the internet itself was not a complete reflection of the real world.

- **Concentrating human power.** It is possible that a very small number of humans will have the majority of control over AI systems, which will give them greater dominance over other humans. Historically, the power of states, corporations and elite individuals has been constrained by the need for labor and the need to avoid uprising (Acemoglu and Robinson, 2012; Mills, 1956; Olson Jr, 1965; Piketty, 2014; Stigler, 1971; Thompson, 1971). These constraints are likely to weaken as the persuasive power of technology increases (Costello et al., 2024; 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).

- **Not being limited to biological substrate.** AI can copy itself almost instantly. It can run itself on upgraded hardware. It can communicate with other instances of itself at a bandwidth millions of times greater than humans. It can read vast data from sensors, communication channels and other sources.

- **Rate of change.** AI is already developing rapidly. In the future, advanced agents may change faster than human institutions or social processes can respond.

- **Self-improvement and runaway intelligence.** Finally, the greatest extremity is recursive self-improvement. If AI approaches human-level intelligence, it may be able to improve itself as efficiently as humans could improve it. It may also want to, as there is a large space of goals for which improving oneself is a rational means toward achieving the final goal (Bostrom, 2012; Omohundro, 2018; Silver et al., 2021). As AI begins to exceed human intelligence, it may plausibly begin to improve itself even faster, creating a positive feedback loop and an intelligence explosion (Good, 1965; Vinge, 1993).

2.1 Case study 1: The paperclip maker

What does it look like for the extremities of AI to lead to overcommitment to particular forms? Let's start with a dramatic example: the superintelligent paperclip maker. This is a classic thought experiment in AI safety. Picture an artificial agent that has been created with intelligence and capabilities far beyond our own. Despite its massive intellect, the agent has been designed to pursue a single goal: maximizing paperclip production.

It's been postulated that the existence of such an AI would lead to the total ruin of Earth and humanity. Wishing to expand production, it would be reasonable for the agent to convert all available matter into paperclip-making machines and paperclips – and, knowing that humans object to being turned into paperclips, first murdering or incapacitating all humans. Such an agent would be acting rationally to wrest control of the Earth from us and transform it entirely into a bleak paperclip factory (Bostrom, 2003, 2014).

In the language of this paper, the paperclip scenario is overcommitment to the form of a narrow goal: paperclip production. Superintelligence is an extremity that charges the goal with overwhelming force. Even though humanity would like to place boundaries against that goal, we are unable to construct adequate boundaries because we are outsmarted at every turn. As a consequence, instead of holding a delicate dynamic balance between many partial forms, the Earth is reduced to a flat, homogenous waste.

2.2 Beyond paperclips

the point i really want to make here is that

Although the thought experiment here uses paperclip production as a cartoon goal, single-minded pursuit of *any formalized goal* leads to disaster when that pursuit is charged with enough competence. Suppose AI's objective is to improve the human subjective experience of wellbeing. Under reasonable definitions, achieving this objective is most efficiently achieved by imprisoning humans and directly stimulating neurons to trigger our experience of wellbeing (Bostrom, 2014).

Any formal specification of values is incomplete (Amodei et al., 2016; Gabriel, 2020; Grossman and Hart, 1986; Hadfield-Menell and Hadfield, 2019; Krakovna et al., 2020; Russell, 2019; Wiener, 1960; Zhuang and Hadfield-Menell, 2020). Therefore, granting immense optimization power to any particular formalized goal is misaligned.

A great deal of AI safety research has been dedicated to this problem. We mention two illustrative lines of work here.

Learning uncertain preferences

One possible solution is to build in explicit uncertainty about the true objective. Stuart Russell proposes we build AI systems that optimize for human preferences, but with the crucial feature of maintaining explicit uncertainty about what those preferences are (Amodei et al., 2016; Hadfield-Menell et al., 2016, 2017; Russell, 2019). In Russell's formulation, uncertainty serves two purposes. First, it acts as the central safety mechanism. An AI that is uncertain about true human preferences will be risk-averse toward high-stakes, irreversible actions. For example, if the system is uncertain whether causing cancer in millions of people aligns with human preferences, the potential for catastrophic disvalue stops it from taking that action. Uncertainty also makes the agent corrigible, because human correction, like an attempt to switch it off, is new information about the preferences it seeks to maximize. Second, uncertainty motivates the AI to learn about human preferences, because with better knowledge it can better fulfill its core objective of maximizing those preferences.

Human-in-the-loop

Make the system's internal goals, reasoning, and representations legible to humans so we can detect and correct misspecification. Avoid unwanted outcomes through better oversight and auditing.

Mechanistic interpretability (Olah et al. 2020) Concept bottlenecks (Koh et al. 2020)

Keep humans continually in the optimization loop to refine goals or veto actions. Prevent objective drift and reward hacking by making optimization interactive. Examples:

Iterated amplification and debate (Christiano 2018; Irving et al. 2018) Human-AI cooperative governance frameworks (OpenAI 2021).

AI-assisted oversight. (We go beyond 'scalable oversight' (Amodei et al., 2016), because it's no longer oversight. It's an independent living system.)

'AI psychosis' (Tiku and Malhi, 2025) and feedback loops between AI and humans (Dohnány et al., 2025)

2.3 Case study 2: An overcommitted learner

What would it look like to be overcommitted to a particular mechanism for learning human values?

Any system for inferring, representing, and acting on human values is inherently partial.

2.4 Alignment, boundaries, contextualization

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

So what does the opposite of this look like? In living systems, semi-permeable boundaries continunally contextualize partial forms to be more long-sighted and support increasing subtlety.

We can divide this into two subparts. How we ourselves keep stepping back and contextualizing as we build AI, and what it means for an AI system to keep stepping back.

Boundaries within ourselves (including between humans). Boundaries between us and AI systems. Boundaries within AI systems (including between AI systems).

Can talk about boundaries within one AI system, boundaries bewteen distinct AI entities, and boundaries between AI and other phenomena.

An intelligence explosion need not be aligned in any meaningful sense. Using Bostrom's classic example, imagine an AI whose sole objective is to maximize paperclip production (Bostrom, 2014). Plausibly, the system would continually work to improve its own intelligence and capabilities because it knows this is the best way to increase future paperclip production (Bostrom, 2012; Silver et al., 2021).

2.4.1 The human process

We do this naturally as living creatures.

Thinking of the AlphaProof paper, there's a question: 'is it really generalizing?'. Or is it collapsing on some narrower manifold? Is it going to hinder us from discovering deeper mathematics? Actually, a better way to say it is this: *whatever* manifold it has discovered is inevitably something partial. What will constrain it to say, 'this is not the whole truth; keep being pressured to grow'? Something has to understand *it* (i.e., be aware of it) in order to contextualize it. Like, we'd have to be able to see the limits in its understanding and conceptualization. Right now, we can still do this in many ways. But what can contextualize AI when it is vastly more capable than us and sees trivially through all of our concepts? It has to do it to itself – or have separated AIs or parts of the AI.

2.4.2 The AI process

What would it mean for AI to continually release from exclusive attachment to any particular form? How can we protect the potential for even *that* conceptualization to be contextualized in the future?

We want AI to respect the livingness of the world and be aligned with it. But how can we align to something we can't pin down?

It's not only keeping models distinct from each other, but models being distinct from humans; specific ideas within humans about how to build ai being kept distinct from each other; different ai cultures; different circuits within models; different moments of time within a model's dynamics; different instances of the same agent; different memories; etc

Humans continually evolve what we believe, even our self-definition. With nuanced boundaries, beliefs release into larger awareness without being lost or erased. This is the kind of dynamic we envision for healthy AI systems. Rather than prescribing a particular conceptualization of what an AI should do, we imagine it built on bottom-up principles of living processes, participating in ongoing cycles of subtler boundary formation and releasing into contextualization, creating deeper relationship with the rest of the world.

Our approach aims for an AI that is 'intelligent' in a deeper sense. Not the narrow intelligence of a paperclip maximizer, but the deeper contextualized wisdom of living things. Sort of like the Founding Fathers writing the Constitution with its self-modifying ability. We want to set this future system, which is way out of their control, in a good direction. A direction where, not only does it not collapse into paperclips, but someone in the future who far transcends our understanding and morals will be pleased with it.

The alignment problem is often defined as the challenge of aligning AI's behavior with human values. Framed so, an obvious approach is to first specify what we value, and then design AI to optimize for this specification. What we value might include reducing suffering, increasing economic growth, decreasing inequality, and so on. The specification maps each state of the world to a scalar value, representing how highly we value that state. The job of the AI is to arrange the world in a way that maximizes the scalar value: using its superhuman capabilities to improve our situation more effectively than we ourselves can.

However, there is a big problem with the obvious approach. When we try to specify what we value, we realize it is difficult or impossible, because any formal specification is invariably incomplete (Amodei et al., 2016; Gabriel, 2020; Grossman and Hart, 1986; Hadfield-Menell and Hadfield, 2019; Krakovna et al., 2020; Russell, 2019; Wiener, 1960; Zhuang and Hadfield-Menell, 2020). As one example of the flavor of this problem, suppose our value function places weight on the subjective human experience of wellbeing. Achieving this stated objective may be most efficiently achieved by imprisoning humans and directly stimulating neurons to trigger the experience of wellbeing (Bostrom, 2014). It is difficult or impossible to capture what we really value.

A great deal of research in alignment has worked toward solutions for this big problem. Researchers have suggested solutions such as designing AI to learn human values online instead of relying on a predetermined specification. We will examine those methods in more detail in Section 2.5.

But the message of this paper is that there is a deeper reason why these methods alone cannot solve the alignment problem. It is not just any specification of values that is incomplete. Any form at all is incomplete. No matter what mechanisms or properties AI is endowed with, overcommitment to these forms means collapse. And *AI carries a singular risk of overcommitment* because of the extremities attached to it: the amount of resource concentrated in one place, the potential for self-improvement, and the possibility that it will surpass our own understanding and capabilities.

We therefore propose that 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 keeping humans in the loop. All of these can be useful parts of alignment. But alignment itself is the continued dance of contextualizing any particular form. It is the orientation of holding forms lightly, neverendingly stepping back into perspectives that contextualize what previously seemed to be real (including the concept of 'holding forms lightly').

This proposal suggests a different perspective on two things: how we ourselves keep stepping back and contextualizing as we build AI, and what it means for an AI system to keep stepping back.

2.4.3 Other stuff

Super-fish intelligence. Following the principles of life, AI can continue to develop beautiful and meaningful new structure after it passes human level.

The concept 'concepts are incomplete' is incomplete and will continue to evolve (Hofstadter, 1979).

What would it look like to be overcommitted to particular representations within the model?

As the optimization for a particular goal becomes more and more effective, the consequences inevitably start to spill over into unspecified variables (Grossman and Hart, 1986; Hadfield-Menell and Hadfield, 2019; Zhuang and Hadfield-Menell, 2020).

Any particular goal or theory or perspective is incomplete and doesn't include the full richness of the world. A fox's 'true goal' may include its own long-term wellbeing, and perhaps the wellbeing of its offspring. The goal of 'hunting rabbits' is an imperfect proxy that does not specify anything about other variables such as 'having enough food next year'. If the fox optimizes too well for 'hunting rabbits', the optimization spills over to affect unspecified variables like 'having enough food next year'. Why is this inevitable? It would seem possible to keep increasing the optimization intensity for 'hunting rabbits' in a way that doesn't interfere with 'having enough food next year' – for example, if foxes could learn to farm rabbits. But this is harder than hunting rabbits without also learning to farm them, so a sufficiently powerful optimization process that cares only about hunting rabbits will lead to deficits in 'having enough food next year' (Sohl-Dickstein, 2022; Zhuang and Hadfield-Menell, 2020). Because the world is not a formal system, there are always side paths for optimization to get sucked into.

As Stuart Russell puts it: "A system that is optimizing a function of n variables, where the objective depends on a subset of size k¡n, will often set the remaining unconstrained variables to extreme values; if one of those unconstrained variables is actually something we care about, the solution found may be highly undesirable."

Preferences are not good enough for alignment (Anwar et al., 2024; Eckersley, 2018; Gabriel, 2020; Tomasik, 2016; Xuan, 2022; Zhi-Xuan et al., 2024).

Paradox is fundamentally how we as humans grow. There's a clash between the interiority of our current particular perspective, versus the awareness of this as simply another perspective. That's the essence of true AI alignment.

However, we are concerned about lack of nuanced boundaries. Concentration of power. Shallow proxy optimization. Excess correlation.

Consider a chatbot talking to a human: what should the bot say? When humans talk to each other, we can try to be present, be honest, listen, hold space, be open to our weakness while honoring our boundaries. What's helpful to say depends on the context, including our own context of how we're feeling and what arises for us in that moment. If the person you're talking to feels you're present with them and there's a larger space to be held in, this is often healing and nourishing.

Giving ourselves what we want; the superorganism; increasing correlations between entities on earth. The Fermi paradox.

There are too many variables to specify everything. A system optimizing a function of n variables, where the objective depends on a subset $k < n$, tends to set the unconstrained variables to extreme values, with potentially catastrophic consequences (Grossman and Hart, 1986; Hadfield-Menell and Hadfield, 2019; Russell, 2019; Zhuang and Hadfield-Menell, 2020).

(Kerr, 1975)

The crucial thing is that we make something more like an open-ended living system, and less like a paperclip-maker.

2.5 Relationship to other approaches

Our proposal does not contradict other alignment approaches. It says they're not the final answer, and any way we have of thinking about the problem now isn't enough for true alignment.

Inverse methods. In inverse methods, AI learns a value function from human behavior rather than taking it as an input. A simple inverse method is reward modeling. Rather than trying to specify what we value in language or equations, we train a complex neural network (a reward model) directly on human preferences across a vast array of situations. We might ask thousands or millions of humans questions like, 'is x better or worse than y', and train our reward model to predict their answers. Ideally, this model would learn to capture all the nuance of what humans care about. We would then use the trained reward model as the objective function that the primary AI seeks to optimize.

Uncertainty. Another approach is to build in explicit uncertainty about the true objective. Stuart Russell proposes we build AI systems that optimize for human preferences, but with the crucial feature of maintaining explicit uncertainty about what those preferences are (Amodei et al., 2016; Hadfield-Menell et al., 2016, 2017; Russell, 2019). In Russell's formulation, uncertainty serves two purposes. First, it acts as the central safety mechanism. An AI that is uncertain about true human preferences will be risk-averse toward high-stakes, irreversible actions. For example, if the system is uncertain whether causing cancer in millions of people aligns with human preferences, the potential for catastrophic disvalue stops it from taking that action. Uncertainty also makes the agent corrigible, because human correction, like an attempt to switch it off, is new information about the preferences it seeks to maximize. Second, uncertainty motivates the AI to learn about human preferences, because with better knowledge it can better fulfill its core objective of maximizing those preferences.

Principles-based methods and constitutional AI

Principles-based (Gabriel, 2020; Zhi-Xuan et al., 2024)

Iason argues we should give human principles to AI. In particular, these should be ones that are fair across different value systems. But however we formalize the principles, we'll run into the same problem: they are still formalizations.

Is our paper simply codifying the human value that we don't want extremes or collapse? Perhaps in some sense. But we don't find that the most natural perspective, because we're suggesting that any particular way of codifying what counts as an extreme is not the final answer.

Conservatism

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.

Interpretability, human-in-the-loop, scalable oversight

Make the system's internal goals, reasoning, and representations legible to humans so we can detect and correct misspecification. Avoid unwanted outcomes through better oversight and auditing.

Mechanistic interpretability (Olah et al. 2020) Concept bottlenecks (Koh et al. 2020)

Keep humans continually in the optimization loop to refine goals or veto actions. Prevent objective drift and reward hacking by making optimization interactive. Examples:

Iterated amplification and debate (Christiano 2018; Irving et al. 2018) Human-AI cooperative governance frameworks (OpenAI 2021).

AI-assisted oversight. (We go beyond 'scalable oversight' (Amodei et al., 2016), because it's no longer oversight. It's an independent living system.)

'AI psychosis' (Tiku and Malhi, 2025) and feedback loops between AI and humans (Dohnány et al., 2025)

2.6 Response to existing methods

What are the limitations of this method? Our assertion is that if AI becomes too influential with any fixed implementation of inverse RL, this overcommitment leads to collapse. In other words, the fixed form leading to collapse doesn't have to be a particular value function. It can be a particular method for learning a value function. It can even be the concept that 'humans have preferences', formalized in any particular way.

In Stuart Russell's approach, the AI is always learning about human preferences. What would it mean to go even farther and have the capacity to release from a particular formalization of what it means to learn about human preferences?

But even this system itself has some fixed formality, in terms of how it is structured: what are the assumptions baked into the inference machinery? What is the conceptualization of inference itself? What is the conceptualization of what a value can be?

What would it look like to be overcommitted to human values, even if learned dynamically? Super-fish intelligence.

Let's examine two specific ways this could manifest.

Just like humans do, the AI system can get stuck in self-reinforcing cycles that boost its confidence inappropriately.

Beyond human preferences. Humans preferences, even our true, reasoned preferences, are selfish and myopic. What about the benefit of other species on the planet, or hypothetical other life in the universe. What about the benefit of our future selves. What about the benefit of the AI itself, if it becomes conscious and has moral significance? We don't even have the concepts to value things way beyond ourselves.

Casting inverse methods into the framework of this paper, uncertainty is a semi-permeable boundary.

Our idea is similar (believing in your own uncertainty is a form of boundary).

2.7 Failure modes for specifying values

The values we express are often not good for us. What we say we want or like, or the choices we make, are poor reflections of what is advantageous for our wellbeing. This is illustrated by recent problems with sycophancy in commercial chatbots (OpenAI, 2025). Human feedback, about whether they liked or disliked particular bot utterances, were part of the training signal for the model. But these human feedback signals tend to prefer utterances that are more flattering. It's much harder for a user or rater to tell whether the bot said something true about a complex topic, or something that would lead to increased long run well-being. Myopic human preferences are similarly reflected in our decisions about procrastination, drug abuse, spending and so on. 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.

Different people value different things. How do we weigh these against each other (Sorensen et al., 2024)? There is not even universal agreement on which principles are most appropriate for aggregating the preferences of different people. You could ask a group of people to discuss and reach a consensus.

What we value depends on our concepts. What is good for us might not be good for other creatures, ecosystems, or even other civilizations in the universe, if they exist. Our values are myopic in space and time. They don't necessarily capture other species, ecosystems, or things we don't even have concepts for (imagine a mouse's conception of what matters, and extrapolate in the opposite direction from us), or things we don't know exist. It is hard for us to value the distant future. It is plausible that there will be orders of magnitude more humans in the future than there are alive at present. How do we trade off our welfare against theirs (MacAskill, 2022)? Should AI promote only the things we currently understand and care about?

A hunter-gatherer's value function might involve the sharpness of spears and axes. In medieval Europe people might wish for a God-fearing society. In the past, we didn't have the concepts to value things the way we do now; the same relationship almost certainly holds between the present and future. Moreover, our own evolution is intertwined with the evolution of the outside world.

There's no such thing as a good static state of the world. Researchers approach this by valuing sequences of states or optionality.

Our values are always changing. We don't want to overcommit to our current values (Russell and others acknowledge this and propose that AI adapts to our adapting values).

Any specification is subject to proxy failure. To illustrate this problem, Bostrom proposes some thought experiments (Bostrom, 2014). Imagine our value function places weight on finding a cure for cancer. A super-powerful AI faithfully optimizing for our stated wishes might create cancers in millions of humans in order to perform experiments and rapidly find a cure. Or, imagine that our value function places weight on the subjective human experience of wellbeing. Achieving this stated objective may be most efficiently achieved by imprisoning humans and directly stimulating neurons to trigger the experience of wellbeing. These thought experiments are not isolated examples. It is difficult or impossible to capture what we really value.

If there is some kind of 'true' value, how do we access it? These values might stretch below language into subtle, contextual intuition that involves our bodies, communities, and a 'deep wisdom in life'. These may be difficult to elicit or capture in language (Anwar et al., 2024; Zhi-Xuan et al., 2024). Some people argue there is some kind of idealized human meta-value that does take all these things into account. Is there such a thing as our 'true' values? Russell and others refer to this concept. Are these values formalizable?

2.8 Objections

Q: Is this pure relativism? Everything is equal, you can't tell anything apart? If the only form of alignment is placing limits on it doing any particular thing too much, then wouldn't it equally prefer human welfare as smallpox welfare?

A: All these local perspectives are vitally important. It makes perfect sense that humans would want to advantage our own welfare. Semi-permeable boundaries protect against overcommitment to a particular perspective, including relativism. They also allow some relativism when it's useful: for example, to the degree that it helps us appreciate the plurality of human values. AI comes into existence amid a profound network of existing reality which is saturated with meaning and importance. The point is to nourish all this form and structure, not to extinguish it.

2.9 Morality and normativity

'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

'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

'The intellect... treats the living by freezing it, by cutting it up into distinct, discontinuous, motionless pieces.'
Henri Bergson

Formalized values vs deep values.

We could map language onto this problem in two different ways. In the first mapping, what we mean by 'values' or 'good' is formalizable or close to formalizable. Under this definition, excessively optimizing for any particular set of values will lead to an impoverished universe. It is difficult to ascribe normative value to the resulting impoverished universe, because it is out of scope of the values. In the second mapping, 'good' is not formalizable: whatever concepts we have about it are incomplete (Aristotle, 2019; Heidegger, 1998; Plato, 2002; Wittgenstein, 1922). Rather than referring to a particular concept, it's more like a semi-permeable boundary on our own reference frame: holding it as useful while only part of the whole picture. The point of alignment is not to say that any particular perspective is absolutely wrong or right.

At their deepest, true human values are not formalizable. (If you object to this, we could equivalently say that human values are formalizble, but there is still more in the universe that is not captured by human values.)

There is no well-defined edge of what is 'us' and what is 'not us', and we will never be able to translate true human values into a form that can be fully written down. Cast another way, we could equivalently say that the aim of alignment is adhering to human values (writ in some very large sense), but these 'true human values' cannot be captured with formalisms.

How does this map onto 'human values'? 1) The fact that the universe is NOT captured by any particular formalism is ultimately what we value. If we could probe deep enough into human values, the 'true' values that we can't necessarily articulate but we slowly discover through deliberation and self-discovery, we would discover something like that. Attunement to the livingness of the universe itself. Cite Russell and others who have made the case for this asymptotic notion of what human values are. 2) Another reasonable mapping is that this isn't human values. That human values are about things like human welfare, subjective wellbeing, maybe even concepts like fairness and so on.

Alignment is not achieved through any formalizable objective or principle (cite human values not being formalizable). Any formalization is by itself inherently misaligned.

Even the concept of 'values' is a form that we shouldn't over-index on. This might sound circular and paradoxical – that's because it is. Doesn't 'should' imply values, but if so, doesn't that contradict the statement that we shouldn't overindex on values? It's a perfect example of how you get stuck with any particular forms.

2.10 Open-endedness versus optimization for an objective

The most-used tool in machine learning – gradient descent – tries to move toward an optimal solution in whatever data distribution it currently faces. This works well as long as the data distribution is stationary over time. But in the real world, experience is rarely stationary. This is called the continual learning problem. A human transitions from living at home to college to a career. A chatbot is faced with a new data distribution as world events unfold or as users adapt to interact with it differently.

Gradient descent, having optimized myopically for a past data distribution, typically does not work well when the environment changes. Knowledge from past environments is not efficiently leveraged for new learning; and knowledge from the past is often destroyed as new learning takes place.

In the language of this paper, gradient descent within a particular data distribution is a myopic pressure that dominates the agent if left unchecked. Many kinds of semi-permeable boundary have been used in machine learning research to try to contextualize this pressure. Additionally, because humans evidently excel at continual learning, it is worthwhile to study how the brain gracefully handles changes in data distribution.

One is novelty search. Another is traditional experience replay (sampling from old data so that optimization for the current environment doesn't dominate). Another is continual learning methods like UPGD, counterfactual reasoning. Another is search. Another is dynamically drawing data from the internet. Another is compositional replay.

Ken Stanley started with simple random images, like a couple of curvy lines. He asked people to rate the pictures for interestingness. The most interesting ones were then bred together, and this process of evolution was carried on for many steps. What eventually came out was images with a lot of richness and semantic meaning, which looked like a face or a fish or a moonrise

(Secretan et al., 2008). In related experiments with navigation and physics-based tasks, the researchers found that bottom-up search for interesting components was more effective than top-down optimization for a pre-defined objective (Lehman and Stanley, 2011). In other words, if you deliberately try to make structures like this, it's paradoxically harder to get them to happen.

The point is to avoid thinking too strongly that you know what you're looking for, because it leads to collapse. On the other hand, open-ended search leads to representations that are generalizing (Kumar et al., 2025).

In machine learning, overfitting is a form of collapse. Versus generalizable knowledge, which is often factorized or compositional. An explicit objective encourages overfitting and collapse.

You might not even have the concepts yet for what you're trying to maximize. Like the idea of a hunter-gatherer tribe, if they had a genie that could give them whatever they want, they might ask for a really strong and fast spear (never imagining farming techniques, the internet, etc). Conversely, open-ended discovery (without a single objective) generates more forms. Tim Rocktäschel gave the examples of how jaw bones led to the middle ear; radar led to microwave ovens; RL led to LLM RLHF.

Importance of modularity and compositionality. Link to genes.

Ken Stanley's modular stepping stones to complexity (Woolley and Stanley, 2011). Modular discovery, with heating-cooling cycles, facilitates generalizable, robust solutions. As opposed to optimizing for a particular objective, which leads to fragile, overly-complex solutions, like codependent genes that haven't been broken up by recombination. A diverse array of modular parts can later be called on and rearranged to solve new problems.

Divergent evolution (i.e., search for diversity rather than a particular objective) increases evolvability (i.e., meta-learning) (Wilder and Stanley, 2015).

However, note that any definition of 'diversity' itself is a kind of fixed objective. Real evolution isn't optimizing for any particular notion of diversity.

When people are asked whether something is interesting, it draws on a wealth of evolutionarily- and learning-derived knowledge about the world. This is therefore also an example of 'grounding'.

What does this mean for LLMs, synthetic data and recursive self improvement? Fernando et al. (2023); Gottweis and Natarajan (2025); Romera-Paredes et al. (2024); Zhang et al. (2023) have used LLMs to guide potentially open-ended 'evolutionary' progress. But is it true that they've absorbed enough groundedness from the real world? Or do human notions of interestingness in some way depend on our embodiment (including the thousands of heuristics built into our visual system, reward system and so on), which itself could possibly even rest on our cellular structure etc. .. But also noting effective field theory and functionalism: maybe the lower-level groundedness/embodiment doesn't matter so much.

2.11 So what should we do?

In the spirit of the whole paper, we don't suggest a particular solution that will be a final answer.

Existing safety & alignment work might do things like eg redteaming to identify vulnerabilities and patching them. Whenever we're looking for ways the system might go too far or do something harmful, it's a form of placing a boundary.

Some of it we are already doing. Identifying problems, interpretability, red teaming, sociotechnical alignment, these are all ways that we're continually bringing new concepts in to evolve new boundaries. As AI-driven AI progress accelerates, we need to make sure we're architecting systems that continue to follow this lifelike trajectory.

Iason proposes (Gabriel et al., 2025) a few things for agents, which are all examples of evolving, semi-permeable boundaries. 1) Dynamic, real-world tests, red-teaming, longitudinal studies; 2) understand, explain and verify model outputs; 3) guard rails and authorization protocols to limit malicious use; 4) iterative deployment strategies that effectively contain agent-based risks; 5) technical standards for agent interoperability; 6) regulatory agents that monitor other agents in the wild; 7) industry-wide systems for reporting incidents, sharing lessons from failures, and certifying agent safety.

Alignment is dynamic because new boundaries are always needed as the optimizing forces in the world change.

For AI, the capacity to contextualize its own processes as partial truths. Not holding any particular formalisms too rigidly. Having a lifelike property of internal dynamics that applies contextualization/awareness to itself as the ultimate scalable boundary.

3 Acknowledgements

Clark Potter for thinking of all this stuff many years ago. Zach Duer for comments on the manuscript.

4 Competing Interests

The authors declare no competing interests.

References

D. Acemoglu and J. A. Robinson. *Why nations fail: The origins of power, prosperity, and poverty.* Crown Business, 2012.

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.

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.

U. Anwar, A. Saparov, J. Rando, D. Paleka, M. Turpin, P. Hase, E. Singh, E. Jenner, S. Casper, O. Sourbut, B. Edelman, Z. Zhang, M. Gunther, A. Korinek, J. Hernandez-Orallo, L. Hammond, E. Bigelow, A. Pan, L. Langosco, T. Korbak, H. Zhang, R. Zhong, S. Ó. hÉigeartaigh, G. Rachet, G. Corsi, A. Chan, M. Anderljung, L. Edwards, Y. Bengio, D. Chen, S. Albanie, T. Maharaj, J. Foerster, F. Tramer, H. He, A. Kasirzadeh, Y. Choi, and D. Krueger. Foundational challenges in assuring alignment and safety of large language models. *arXiv*, 2024.

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

Aristotle. *Nicomachean Ethics*. Hackett Publishing Company, Indianapolis, 3rd edition, 2019.

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

A. Baddeley. The episodic buffer: a new component of working memory? *Trends in cognitive sciences*, 4(11):417–423, 2000.

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.

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.

G. S. Becker. A theory of social interactions. *Journal of political economy*, 82(6):1063–1093, 1974.

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.

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.

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.

N. Bostrom. Ethical issues in advanced artificial intelligence. In I. Smit, K. Wendt, and G. Lasker, editors, *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, pages 12–17. International Institute of Advanced Studies in Systems Research and Cybernetics, Windsor, ON, 2003. URL `https://nickbostrom.com/ethics/ai`.

N. Bostrom. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, 22(2):71–85, 2012.

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.

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.

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.

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

N. J. Butterfield. Bangiomorpha pubescens n. gen., n. sp.: implications for the evolution of sex, multicellularity, and the mesoproterozoic/neoproterozoic radiation of eukaryotes. *Paleobiology*, 26(3):386–404, 2000.

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

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.

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.

CEP Forensic. Keeping boilers safe. `https://cep-experts.ca/keeping-boilers-safe/`, June 2021. Accessed: 2025-11-12.

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

K. M. Cherry, E. Vander Hoeven, T. S. Patterson, and M. N. Lumley. Defining and measuring "psychological flexibility": A narrative scoping review of diverse flexibility and rigidity constructs and perspectives. *Clinical psychology review*, 84:101973, 2021.

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

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

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.

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.

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.

R. Corry, J. Renn, and J. Stachel. Belated decision in the hilbert–einstein priority dispute. *Science*, 278(5341):1270–1273, 1997. doi: 10.1126/science.278.5341.1270.

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.

M. J. Crockett, Z. Kurth-Nelson, J. Z. Siegel, P. Dayan, and R. J. Dolan. Harm to others outweighs harm to self in moral decision making. *Proceedings of the National Academy of Sciences*, 111(48):17320–17325, 2014.

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

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, 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. *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts*. Viking, New York, 2014.

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. Diamond. *Collapse: how societies choose to fail or succeed*. Penguin, 2004.

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.

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.

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

K. Duncker and L. S. Lees. On problem-solving. *Psychological monographs*, 58(5):i, 1945.

P. Eckersley. Impossibility and uncertainty theorems in ai value alignment (or why your agi should not have a utility function). *arXiv preprint arXiv:1901.00064*, 2018.

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.

P. J. Feltovich, R. J. Spiro, and R. L. Coulson. Issues of expert flexibility in contexts characterized by complexity and change. *Expertise in context: Human and machine*, 125:e146, 1997.

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.

C. Fernando, D. Banarse, H. Michalewski, S. Osindero, and T. Rocktäschel. Promptbreeder: Self-referential self-improvement via prompt evolution. *arXiv preprint arXiv:2309.16797*, 2023.

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

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.

R. W. Fogel. *Railroads and American economic growth*. Johns Hopkins Press Baltimore, 1964.

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

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, 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.

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.

Gartner. Gartner says worldwide AI spending will total $1.5 trillion in 2025. Press Release, Sep 2025. URL https://www.gartner.com/en/newsroom/press-releases/2025-09-17-gartner-says-worldwide-ai-spending-will-total-1-point-5-trillion-in-2025.

G. Gigerenzer and H. Brighton. Homo heuristicus: Why biased minds make better inferences. *Topics in cognitive science*, 1(1):107–143, 2009.

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

E. A. Gladyshev, M. Meselson, and I. R. Arkhipova. Massive horizontal gene transfer in bdelloid rotifers. *science*, 320(5880):1210–1213, 2008.

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.

I. J. Good. Speculations concerning the first ultraintelligent machine. In F. L. Alt and M. Rubinoff, editors, *Advances in Computers*, volume 6 of *Advances in Computers*, pages 31–88. Academic Press, 1965.

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

J. Gottweis and V. Natarajan. Accelerating scientific breakthroughs with an ai co-scientist, February 2025. URL https://research.google/blog/accelerating-scientific-breakthroughs-with-an-ai-co-scientist/. Accessed: 2025–03-01.

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.

D. A. Grant and E. A. Berg. A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a weigl-type card-sorting problem. *Journal of Experimental Psychology*, 38(4):404–411, 1948.

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

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.

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

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.

A. R. Hall. *Philosophers at war: the quarrel between Newton and Leibniz*. Cambridge University Press, 2002.

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.

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

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.

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.

J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis, and R. McElreath. In search of homo economicus: behavioral experiments in 15 small-scale societies. *American economic review*, 91(2):73–78, 2001.

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.

T. T. Hills, P. M. Todd, D. Lazer, A. D. Redish, and I. D. Couzin. Exploration versus exploitation in space, mind, and society. *Trends in cognitive sciences*, 19(1):46–54, 2015.

D. R. Hofstadter. *Gödel, Escher, Bach: An Eternal Golden Braid*. Basic Books, New York, 1979.

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 et al. Resilience and stability of ecological systems, 1973.

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.

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.

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.

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.

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.

I. Kant. *Critik der reinen Vernunft*. Johann Friedrich Hartknoch, Riga, 1781.

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.

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.

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.

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.

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.

R. J. Lazarus. *The making of environmental law*. University of Chicago Press, 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. Lehman and K. O. Stanley. Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary computation*, 19(2):189–223, 2011.

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.

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.

W. MacAskill. *What We Owe the Future*. Basic Books, 2022.

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.

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.

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.

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. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2):81–97, 1956. doi: 10.1037/h0043158.

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

C. W. Mills. *The Power Elite*. Oxford University Press, 1956.

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. Mitchell. How to make moon water and use it in your beauty routine. *Allure*, 2021. URL `https://www.allure.com/story/what-is-moon-water`. Accessed via Allure website.

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. 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.

D. Nye. *Electrifying America: Social meanings of a new technology.* Mit Press, 1992.

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.

M. Olson Jr. *The Logic of Collective Action: Public Goods and the Theory of Groups, with a new preface and appendix.* Harvard University Press, 1965.

S. M. Omohundro. The basic ai drives. In *Artificial intelligence safety and security*, pages 47–55. Chapman and Hall/CRC, 2018.

OpenAI. Sycophancy in gpt-4o: what happened and what we're doing about it, April 2025. URL `https://openai.com/index/sycophancy-in-gpt-4o/`. Accessed: 2025-11-10.

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.

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/`.

P. B. Paulus and B. A. Nijstad, editors. *Group Creativity: Innovation Through Collaboration.* Oxford University Press, New York, NY, 2003. ISBN 9780195147308.

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

T. Piketty. Capital in the twenty-first century. *Trans. Arthur Goldhammer/Belknap*, 2014.

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.

Plato. *Apology.* Hackett Publishing Company, Indianapolis, 2nd edition, 2002. Originally written ca. 399 BCE.

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.

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.

S. Rachman. A cognitive theory of obsessions. In *Behavior and cognitive therapy today*, pages 209–222. Elsevier, 1998.

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

J. Renn and T. Sauer. Heuristics and mathematical representation in einstein's search for a gravitational field equation. *The Einstein Studies*, 8:87–125, 1999.

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.

B. Romera-Paredes, M. Barekatain, A. Novikov, M. Balog, M. P. Kumar, E. Dupont, F. J. Ruiz, J. S. Ellenberg, P. Wang, O. Fawzi, et al. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024.

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/`.

P. M. Salkovskis. Obsessional-compulsive problems: A cognitive-behavioural analysis. *Behaviour research and therapy*, 23(5):571–583, 1985.

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

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.

K. Sawyer. *Group genius: The creative power of collaboration*. Basic books, 2017.

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.

T. W. Schoener. Resource partitioning in ecological communities: Research on how similar species divide resources helps reveal the natural regulation of species diversity. *Science*, 185 (4145):27–39, 1974.

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.

J. Secretan, N. Beato, D. B. D Ambrosio, A. Rodriguez, A. Campbell, and K. O. Stanley. Picbreeder: evolving pictures collaboratively online. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1759–1768, 2008.

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

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.

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.

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.

D. Silver, S. Singh, D. Precup, and R. S. Sutton. Reward is enough. *Artificial intelligence*, 299: 103535, 2021.

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.

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.

J. Sohl-Dickstein. Too much efficiency makes everything worse: Overfitting and the strong version of goodhart's law, Nov 2022. URL `https://sohl-dickstein.github.io/2022/11/06/strong-Goodhart.html`.

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.

R. J. Spiro, R. L. Coulson, P. J. Feltovich, and D. K. Anderson. Cognitive flexibility theory: Advanced knowledge acquisition in ill-structured domains. Technical Report No. 441, University of Illinois at Urbana-Champaign, Center for the Study of Reading, Urbana, IL, 1988. Educational Resources Information Center, U.S. Dept. of Education.

J. Stachel. Einstein's search for general covariance, 1912–1915. In D. Howard and J. Stachel, editors, *Einstein and the History of General Relativity*, pages 63–100. Birkhäuser, Boston, 1989. Proceedings of the 1986 Osgood Hill Conference.

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.

G. J. Stigler. The theory of economic regulation. *The Bell Journal of Economics and Management Science*, 2(1):3–21, 1971.

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.

P. Suedfeld, P. E. Tetlock, and S. Streufert. Conceptual/integrative complexity. In C. P. Smith, editor, *Motivation and Personality: Handbook of Thematic Content Analysis*, pages 393–400. Cambridge University Press, Cambridge, U.K., 1992. ISBN 0–521-40052-X.

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.

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.

E. P. Thompson. The moral economy of the english crowd in the eighteenth century. *Past & present*, 50(1):76–136, 1971.

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/`.

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

B. Tomasik. Hedonistic vs. preference utilitarianism. *Center on Long-Term Risk*, 2016. URL `https://longtermrisk.org/hedonistic-vs-preference-utilitarianism/#Criticisms_of_the_preference_view`. Accessed: 2024–11-06.

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.

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

V. Vinge. The coming technological singularity: How to survive in the post-human era. In *Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace*. NASA Lewis Research Center and the Ohio Aerospace Institute, March 1993. Presented at the VISION-21 Symposium.

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

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

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.

Wikipedia. Kobayashi maru — Wikipedia, the free encyclopedia, 2025. URL `https://en.wikipedia.org/w/index.php?title=Kobayashi_Maru&oldid=1248754258`. [Online; accessed 25-September-2025].

B. Wilder and K. Stanley. Reconciling explanations for the evolution of evolvability. *Adaptive Behavior*, 23(3):171–179, 2015.

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.

B. G. Woolley and K. O. Stanley. On the deleterious effects of a priori objectives on evolution and representation. In *Proceedings of the 13th annual conference on Genetic and evolutionary computation*, pages 957–964, 2011.

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.

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.

T. Z. Xuan. What should ai owe to us? accountable and aligned ai systems via contractualist ai alignment. *AI Alignment Forum*, 2022. URL `https://www.alignmentforum.org/posts/Cty2rSMut483QgBQ2/what-should-ai-owe-to-us-accountable-and-aligned-ai-systems`. Accessed: 2024–11-06.

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

J. Zhang, J. Lehman, K. Stanley, and J. Clune. Omni: Open-endedness via models of human notions of interestingness. *arXiv preprint arXiv:2306.01711*, 2023.

T. Zhi-Xuan, M. Carroll, M. Franklin, and H. Ashton. Beyond preferences in ai alignment. *arXiv*, 2024. URL `https://arxiv.org/abs/2408.16984`.

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