Alignment, Boundaries, Contextualization

October 27, 2025

Abstract

The alignment problem in artificial intelligence (AI) is often posed as the problem of building AI to act in accordance with human values or principles. Here, we suggest a different way to think about alignment. True alignment is a rich, ever-changing, imperfect array of semi-permeable boundaries that contextualize myopic forms. Contextualization nuances the functioning of systems and builds potential for unexpected phenomena to spontaneously arise and evolve in a life-like way.

Most alignment research is premised on the notion of aligning AI with some well-designed goals or principles. In this paper, we take a different view. We argue that true alignment is not achieved through any formalizable objective or principle. In fact, any fixed conceptualization, even one incorporating uncertainty, is by itself inherently not aligned.

Each form – an equation, a gene, a concept, an institution – is myopic in the sense that it is only a small aspect of the world. But the existence of many partial forms, maintaining separate identities while dynamically working together at many scales, constitutes the rich livingness of the world. Collapse is a failure mode where forms blend together or a particular form gains excess traction, resulting in a loss of functional diversity. Boundaries protect against collapse. Without sufficient boundaries, cancer spreads, concentrated power reduces human welfare, invasive species choke out complex ecosystems.

AI charges particular myopic forms with enormous leverage. Humanity is presently diverting an unprecedented volume of energy, information and computation into AI training and inference. If the training objective or the form of the model at inference time remains fixed while the system gains increasing resources 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 to support ongoing deepening of the world's organization.

In the following sections, we use examples from natural and human systems to illustrate how boundaries place intelligent constraints that limit collapse. Crucially, boundaries must be semi-permeable. By maintaining distinctiveness of separate forms while also allowing them to interact modularly, semi-permeable boundaries contextualize myopic forms as part of larger systems. In the resulting delicate dynamic balance, there are many possibilities for new phenomena. In this light, we recast alignment as an evolving network of semi-permeable boundaries that contextualize any particular form of AI to support subtle functioning and ongoing arising of new, unknown structure.

1 Cell membranes and semi-permeable boundaries

Boundaries, collapse and contextualization are abstract ideas. To make them more concrete, we will walk through a series of examples from living systems. Within each example, the ideas are commonsense and straightforward. Our hope is that by looking at multiple examples it stands out how general and fundamental they are.

Without boundaries, interactions cause collapse, which is a reduction in meaningful diversity. The cell membrane is a key boundary of living systems. It 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, which 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).

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, the process of enveloping, 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.

A phenomenon we will see repeatedly across systems is that 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 concept will be central for the alignment problem, which we return to in detail in the last section of the paper.

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

2 Diversity and problem solving in groups

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

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

However, 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; Flowers, 1977; Frey and Van de Rijt, 2021; Janis, 1972; Stasser and Titus, 1985). In the words of mathematician Jean Écalle, "I could also observe, time and again, how too deep an immersion in the math literature tended to stifle creativity" (Wood, 2023). 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.

There are multiple types of effective semi-permeable boundary in groups, including: creating decentralized topologies where group members only communicate with nearby neighbors (Becker et al., 2017); 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); and promoting leadership styles where one person's views are less likely to dominate (Flowers, 1977; Leana, 1985). In a later section, we will look at boundaries within an individual, such as skepticism, that make it easier to receive information 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. The harmful effect of premature communication 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.

Historians of science have presented Einstein's path to general relativity as a case study in persevering with one's own independent approach to a problem (Corry et al., 1997; Renn and Sauer, 1999; Stachel, 1989). In the years leading up to 1915, Einstein was guided by deep physical insight but struggled to express his ideas in mathematical language. Some scholars argue this struggle was essential. If Einstein had relied on the powerful mathematical techniques being developed by David Hilbert and others, he may have arrived at a solution that worked superficially but lacked deeper physical meaning.

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.

3 Genes and modularity

The "problem" is collapse: excess formalism/correlation. In asexually reproducing species, this is a single assembly of genes being rigidly and permanently stuck together. Boundaries contextualize that formalism, still allowing it to exist to a degree, but it doesn't have the final say anymore. Sex still allows assemblies, but now they're not absolute/permanent. By having a set of distinct (not too correlated) units, they can work together dynamically in rich ways.

In asexual reproduction, all the genes are stuck together rigidly, and that assembly is doomed to its hard-coded form for the whole lineage forever. It's like if you were forced to only think one thought forever. It's a collapse out of dynamism, pinning the genome of the lineage down to one structural form. Sexual reproduction is a boundary that contextualizes those formalisms. The formalism of the genome can still exist temporarily (eg the lifetime of one organism), and aspect of it that work well are propagated. But if it doesn't work, it's discarded and another one is tested – like how thoughts are always evolving and being tested against each other in your brain.

That rigidity creates two problems. First, you don't generate flexible new possibilities (no recombination). Second, even if a particular assembly is good now, it's so isolated from the dynamism of the world that as it inevitably slowly decays, it's stuck in that decay (mutations accumulate in a lineage). At a species level, this manifests as Muller's ratchet.

Replicators are selected for being good at replicating. But short-sighted optimization for immediate replication leads to long-term trouble (Dawkins, 1976; Traulsen and Nowak, 2006; Wagner and Altenberg, 1996; Wilson, 1980). For example, pathogens perform better in the short-term if they voraciously consume the host. But if they kill their hosts too quickly, the infection won't spread in the population, and the pathogens will eventually perish. Bacteria have evolved mechanisms to recognize and suppress defectors among themselves (West et al., 2006). These mechanisms are boundaries that contextualize the myopic interests of individual pathogens, allowing the pathogen community as a whole to prosper in the longer term.

At a larger scale, sexual reproduction contextualizes the myopic force of natural selection. Nearly all eukaryotes reproduce sexually, even though sex is a costly strategy (Goodenough and Heitman, 2014; Lehtonen et al., 2012; Maynard Smith, 1971, 1978; Speijer et al., 2015). For a sexual species, half of all organisms don't have the ability to reproduce, which should harshly curtail population growth and allow asexual competitors to dominate. Moreover, finding a mate is not easy in a vast and dangerous world. So what are the evolutionary benefits of sex?

All organisms are subject to periodic mutations in their genetic code, due to copying errors and damage, and the vast majority of mutations are either neutral or harmful to fitness. In an asexually reproducing species, whenever a creature experiences a germline mutation, that creature's entire lineage is stuck with the mutation forever. (The odds of the mutation being spontaneously reversed by another mutation are vanishingly small.) So within each lineage, the mutation load only increases with time. When, by chance, the least-loaded lineage dies off, the maximum fitness of the species is irreversibly lowered. This is Muller's ratchet: the loss of a less-burdened lineage is like a ratchet click that the species can never recover from (Keightley and Otto, 2006; Muller, 1932). Thus, the overall fitness of an asexual species experiences steady downward pressure from increasing mutation load. With asexual reproduction, the whole lineage suffers the weight of any mutations – like groupthink where a whole group has the same wrong idea and there's nothing to challenge it.

In a sexual species, by contrast, harmful mutations can be reversed. Suppose creature X has a mutation and creature Y has a different mutation. X and Y mate and produce offspring. Some offspring will have both mutations, some will have one but not the other, and crucially, some will have neither. With sexual reproduction, it is possible to turn the ratchet backward and produce offspring that are fitter than either of the parents. Sex was a critical innovation and deeply intertwined with the emergence of complex multicellular life (Butterfield, 2000).

Sex is a semi-permeable boundary. In a sexual species, the occasional beneficial mutation can still be utilized, but harmful mutations are recombined out of the population. By reducing selfing, sex creates a space where the existing diversity and dynamism of the underlying substrate discovers new interesting solutions, re-routing the energy of short-sighted optimization in a longer-sighted direction.

Breaking up co-adaptation.

By making things modular, boundaries encourage each entity to have interactions with many other entities. When a system is under pressure, modularity makes each entity become more 'agentic': it has to develop a wisdom that's both independent and inter-functional.

In general, evolution continues discovering boundaries to contextualize particular myopic forms. Boundaries allow multiple experiments to run in parallel without overwriting each other – preventing the space of forms from being collapsed.

Boundaries also protect individual genes from being blended away. If computers were analog rather than digital, it would be very hard to do error correction because noise would be additive. Likewise, discretiziation avoids genes getting confused with graded copies of themselves. That's what would happen with continuously blended inheritance.

https://chatgpt.com/c/68ec4a38-4dc0-8325-888c-106db5b85a08

4 Ecosystems

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

Prior to the arrival of Europeans, the gray wolf was an apex predator in the region of the Rocky Mountains now making up 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 dominated to excess. Predation supplied a semi-permeable boundary:

it placed 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 contain a large and finely-tuned array of semi-permeable boundaries, including predation and parasitism, competition, resource limitation, 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 self-centered motives of each species, when contextualized by semi-permeable boundaries, work toward open-ended enriching of life.

Human drives especially are often 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). It's again a paradox – we try to achieve certain aims for our own benefit (like resource extraction), yet if we're too successful in those particular aims, the end result is negative for our overall welfare.

Fortunately, there are some boundaries on human ecosystem impacts. One boundary 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. However, humans also create our own boundaries, including state regulation, self-policing and market-driven forces like environmental certifications. Through the long-sightedness of our own intelligence, we sometimes foresee the consequences of excess extraction and place limits on it. These 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).

Boundaries create the space in which spontaneous life-like unfolding can happen. If the human drive for extraction is pursued too single-mindedly, it collapses the subtle complexity of nature. But if that human drive is bounded and rich nature and wilderness remain, then we have a reservoir of energy within which new things are bubbling up. (cite effects of wilderness on humans; also ongoing evolution in nature; also discovery of drugs and science from wilderness (which includes outer space))

Finally, 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 often leave some destruction in our wake, yet the extraction fuels explosion of technology, art, music, and human experience. This is another principle that recurs across many domains of life.

By allowing formalism at the small scale (i.e., by allowing entities to become more distinctly themselves), boundaries make things compositional, as opposed to being overfit and entangled. Organisms of different species can run around and interact with the world modularly. On many levels. The gene, the individual organism, the species. Imagine if there was only one organism on Earth. Say, a big tree that grew an intricate root and canopy system over the whole planet, with each part highly specialized to function exactly with its very specific neighbors. It's the fact that the real world is NOT like that; that it's separated organisms, which creates the possibility for combinatorially many kinds of interactions and the explosion of rich activity that results.

Also, notice that when entities are agentic, like organisms in an ecosystem, there's a new kind of collapsing pressure. Rather than interactions just passively causing collapse by increasing correlations (like diffusion), an agent can actively impose its will on other entities (eg an animal eating another animal).

5 Art: boundaries support spontaneous novelty

The best art escapes simple definition. It encourages many interpretations and metaphors. People find meaning in it that's intimately connected to the uniqueness of their own lives. With some great art these meanings continue to evolve even across centuries. For Kant, 'aesthetic ideas', unlike 'concepts', could not be adequately expressed in words; aesthetic ideas gave rise to endless interpretive play (Kant, 1781). Schiller similarly emphasized that the meaning in art 'unfolds in freedom': it is not pinned down to one definition only (Schiller, 1795). Eco agrees that great art is 'open', offering multiple interpretations, and he goes a step farther to argue that openness should not be confused with vagueness or randomness. A well-crafted work invites interpretive possibilities rather than collapsing into chaos or relativism (Eco, 1962).

Like with the other living systems we've explored, our experience of art becomes impoverished if a single interpretation or metaphor dominates at the expense of others. A physical artwork – the oil on a canvas, the notes on a staff, the words on a page – can be one of the most extraordinary forms of semi-permeable boundary in the living world. An artist wraps the subtlety of their experience into a form that interacts with the viewer. When an individual or a society starts to interpret it, the pressures to try to reduce it to a particular understanding are resisted by the structure of the art itself (Barthes, 1970; Empson, 1930; Gadamer, 1960; Wimsatt and Beardsley, 1946). The consequence is a proliferation of new depth and internal meaning.

6 Laws: an evolving patchwork

Individual actors in a society and in an economy each act from their own perspective. This perspective is not always selfish in the sense of maximizing wealth or physical wellbeing for the actor (Becker, 1974; Crockett et al., 2014; Henrich et al., 2001), but it is always myopic because no individual knows everything or fully understands the motives and beliefs of others.

Without boundaries, a particular agent's perspective can dominate, resulting in collapse and an impoverished system. For example, companies' 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).

Law is a boundary against dominance of any particular agent's drive. Laws hold in check the drives of individual entities. Someone 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 therefore 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).

Despite its adaptiveness, law has historically not always been an effective boundary for checking the power of the most elite individuals and most successful businesses. Instead, their power has been limited by factors like competition with other self-interested entities; and also 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). There is a possibility that these constraints will weaken in the future. In some forecasts AI and robotics may begin to supplant human labor. Technology also brings increased persuasive power (Costello et al., 2024; Woolley and Howard, 2018) and the potential for avoiding uprising by more effectively influencing beliefs.

6.1 Interpersonal dynamics

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.

In the framework of this paper, there are many different perspectives and motives in the world. Without boundaries, interactions tend to result in one being dominated by another: a patient's own beliefs may be replaced with those of an analyst, or the desires of one person in a relationship might be ignored. With semi-permeable boundaries, each agent's autonomy is supported but also contextualized as part of a broader system. New forms then spontaneously emerge: mutual understandings, relationships, communities, cultures.

It's interesting to note that at a smaller scale, boundaries increase formalization: they allow separate entities to have distinctively their own forms. Interpersonal boundaries support separate identities existing and differentiating. But at a larger scale, boundaries resist formalization because multiple forms simultaneously develop their own contradictory perspectives. Because there are multiple people with their own distinct ideas, the larger system becomes less formal (or at least requires a richer formal description).

6.2 Information in the brain

The brain is somewhat miraculous in that it manages to keep so many representations 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. There are many more; the brain is the most extraordinary example in nature of a system of semi-permeable boundaries supporting the proliferation of multitudinous forms that are kept distinct yet 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 maintain distinct neural representations. Lateral inhibition was first studied in the visual system, 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. 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).

Global inhibition also supports the existence of distinct forms. In the hippocampal formation and connected areas, some cells are tuned to particular directions the animal's head could be facing. Inhibition creates a winner-take-all effect, integrating over intermittent noisy evidence (like vestibular signals when the head turns) to create a single stable representation of the head direction (Rolls, 2022; Zhang, 1996). Inhibition prevents the signals in some channels from getting blended or overwritten by the signals in other channels.

The brain uses inhibition organized into oscillatory dynamics to keep representations separated (Jensen and Mazaheri, 2010; Klimesch et al., 2007; Lisman and Jensen, 2013; Roux and Uhlhaas, 2014). Distinct memory 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 items in phase space, multiple items are held simultaneously without interference.

More broadly, healthy cortical 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 near-critical regime, the brain has access to a huge repertoire of patterns that it can explore temporarily without getting stuck in any one state. 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). More extreme hypersynchrony leads to severe dysfunction. For example, in Parkinson's disease, basal ganglia and cortical circuits collapse into excess synchrony and lose the sensitivity needed to flexibly guide nuanced motor outputs (Brown, 2003; Hammond et al., 2007). In healthy, the brain maintains near-criticality by constantly homeostatically fine-tuning the balance between excitation and inhibition (Cocchi et al., 2017; Haldeman and Beggs, 2005; Shew et al., 2011).

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.

Finally, the human brain tends to discretize the world 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 ability for reasoning (Chomsky, 1957; Fodor, 1975; Kurth-Nelson et al., 2023; Lake et al., 2015; Pinker, 1994).

Again, boundaries increase formality at a local scale. There is less blending or overwriting between entities, so each entity has more distinctively its own form. At the same time, at a larger scale, boundaries avoid excess formality by preserving a space of distinct, contradictory forms within the larger system.

6.3 Drives, goals and impulsivity

Motivation functions best in support of the organism's overall health when it doesn't excessively collapse into particular drives, strategies, or goals, but instead maintains broader flexibility.

Animals balance multiple innate drives, like nutrition, osmotic balance, temperature regulation, avoiding illness and injury, and reproduction (Saper and Lowell, 2014; Schulkin and Sterling, 2019; Sewards and Sewards, 2003). Each drive is a proxy for evolutionary fitness, yet excess optimization for one drive alone decreases fitness (John et al., 2023; Kurth-Nelson et al., 2024). For example, if calorie intake is maximized to excess, without being balanced by other drives, 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). In general, if a single drive dominates, the organism becomes unwell.

Paradoxically, a particular strategy for optimizing a drive can even dominate at the expense of achieving the aims of the drive. 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 space of low-level 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. Unbalanced optimization 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. Excessive optimization for narrow goals happens at the expense of a broader balance of goals, and at the expense of the health of the organism or others. As a general rule of thumb, a single form dominating to excess is the opposite of health.

A broad class of boundaries on particular drives, strategies, goals is *cognitive control* (Botvinick et al., 2001; Braver, 2012; Miller and Cohen, 2001; Miyake et al., 2000). In the case of overeating, control overrides the food-seeking drive. In the case of the chickens, control overrides 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.

When nothing stops a myopic drive or strategy from dominating behavior, it tends to achieve its aims through a shortest-path mechanism. The simple strategy of "approach food", if not contextualized by cognitive control or other mechanisms, achieves its aims of moving toward the food, without accomplishing the deeper goal for which moving toward food is simply a proxy. When the chicken runs directly toward the food, it thus spends energy without achieving the deeper goal. This energy is discharged or short-circuited. Boundaries, on the other hand, translate the pressure of a drive into higher-order structure. Symmetry is broken because the best way to approach the food is no longer the shortest-path in primitive space. This effect – boundaries constraining a proxy and thereby leading to the deeper drive being expressed in a more structured way – is another pattern we'll encounter across multiple living systems. As long as boundaries limit the dominance of particular entities, new living structure continues to spontaneously emerge.

6.4 Frames and perspectives

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 is 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 computer to change the simulated Klingons from belligerent to helpful (Wikipedia, 2025), thus rescuing the crew and avoiding war.

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 (De Bono, 1970; Duncker, 1945; Goffman, 1974; Javed and Sutton, 2024; Korzybski, 1933; Lakoff and Johnson, 1980; Ohlsson, 1992; Popper, 1934; Saffo, 2008). 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. Switching between them is advantageous for flexible problem solving and deeper understanding.

In Karl Duncker's famous Candle Problem, research participants were given a candle, a box of thumbtacks, and some matches, and they were challenged to attach the candle to a vertical corkboard in such a way that it could burn without dripping wax onto the table below (Duncker and Lees, 1945). Most participants tried to either tack the candle directly to the corkboard, or melt the candle onto the corkboard, but neither solution was stable. Successful participants emptied the tacks out of their box, tacked the box itself to the corkboard, and then secured the candle in the box. Success required escaping what Duncker called 'functional fixedness' – excess attachment to one perspective. There's no reason why the box can't be conceptualized as a candle holder; shifting to this different framing solves the problem.

Surprisingly, asking the same person to solve the same problem twice – and then aggregating the two solutions – can boost performance. This has been called 'wisdom of inner crowds' (Herzog and Hertwig, 2009, 2014; Rauhut and Lorenz, 2011; Stroop, 1932; Vul and Pashler, 2008). If the solution is elicited twice in quick succession, the benefits are typically minimal compared to asking once: the two answers tend to be very similar. But if the person returns to the problem later and thinks about it afresh, then averaging their two answers carries a substantial advantage. Interestingly, the same benefit can be obtained without a time gap by explicitly requesting that the participant take a different approach for their second solution, without giving any additional information about the problem.

The capacity to adopt multiple perspectives is, fittingly, described in several different ways across different areas of psychology and cognitive science. In clinical psychology, '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). People with 'adaptive expertise' are those who use multiple distinct interpretations, analogies or schemas to perform better in complex problem domains. Adaptive experts dynamically evaluate the appropriateness of different approaches (Feltovich et al., 1997; Hatano and Inagaki, 1984; Spiro et al., 1988). 'Integrative complexity' is the cognitive ability to first differentiate multiple perspectives on a problem domain and then identify conceptual connections between these distinct perspectives (Suedfeld et al., 1992; Tetlock, 1986). Gigerenzer argues that humans deploy many different 'heuristics', each of which only processes a problem through a narrow lens. By contextually switching between heuristics, we get good-enough solutions to many problems (Gigerenzer and Brighton, 2009). 'Set shifting' is the ability to transition between different 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 dominates to the detriment of healthy functioning. A recent case study reported a 47 year old man in India who could not stop thinking about how he might have swallowed a plate, a tin sheet, or even a building, leading to significant distress (Karnam et al., 2025). 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 zooming in on one hypothesis or framework.

In the absence of obsessions and delusions, we might momentarily entertain a strange notion, but it quickly disappears and then we think about something else. Or perhaps it becomes the seed for a creative idea as it interweaves with other evidence and beliefs. Under ideal conditions, our thinking is responsive to changes in context and adapts as we exhaust the utility of pursuing a particular direction. Here again is the principle that if boundaries contextualize myopic forces, then the new subtlety is spontaneously expressed.

In a more lighthearted 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. Just like with the chickens in the previous section, overriding this logic requires a higher-order correction term, which yields a closer approximation to the truth. The correction term is relatively subtle: tides arise from differential tugging over long distances in a body of water that is free to slosh around. Subtle correction terms are often hard-won knowledge originating from thoughtful interactions with the world. But we only profit from those interactions if we maintain skepticism of our own existing beliefs. Skepticism is a boundary holding our existing beliefs in check, so they do not dominate our ideas forever.

Despite all these examples, the existence of narrow points of view is not a problem. In fact, it's necessary. Even obsession can be powerful when we obsess on a problem at work and sometimes achieve great results. A delusion-like framework can seed a scientific revolution. The point is not to shut down all particular drives or concepts. The point is to limit them from becoming the sole and absolute determinants of behavior. For example, I might work obsessively on a project, while also having a rule that I must go to bed at 10 pm. Therefore, the most useful boundaries are semi-permeable. They don't block us from taking a strong perspective, but they do 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.

In the Lisa Stardust example, boundaries also protect the mystic beliefs. Which is exactly what you want. The mystic frame should be free to explore its own reality too. Stardust doesn't know a priori that the mystic frame is wrong. It could be like heliocentrism where you have to start by believing something that everyone else thinks is crazy. But more crucially, even once she discovers that the mystic frame doesn't do a good job 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. Contextualization is the wisdom that you can hold the mystic frame for what it is, while simultaneously understanding that the newtonian explanation is better for launching rockets or something.

I listened to an interview where someone said that our physics and biology don't address the problem of minds. I saw that I immediately jumped to a thought that's very familiar in myself: 'physics does address minds because physics gives rise to chemistry which gives rise to biology which gives rise to behavior; if we could run a big enough physics simulation it would produce minds'. I then kind of shelved the rest of what he was saying because I had *explained away* his perspective. Another example is when we explain away the beliefs of people with different political views. In the United States, Democrats will often say that Republicans have their views 'because they didn't get enough education'. Republicans sometimes say Democrats have their views 'because they are stuck in a bubble'. Then we don't engage with the real contents of the views within ourselves. Its a way that our belief-system asserts dominance (collapse), not allowing diversity of perspective. The semi-permeable boundary is awareness of the explaining-away thought without being exclusively identified with it: contextualizing that thought, holding it lightly.

7 AI alignment

Across a bunch of living systems, we've looked at how semi-permeable boundaries place selective limits on interactions so that, instead of blending or overwriting each other, forms maintain distinctiveness while functioning as part of larger systems. With these intelligent protections, new structure is free to continue naturally emerging in a life-like way. Now we apply the same lens to the AI alignment problem.

7.1 The alignment problem

There is no single, universally agreed definition of the alignment problem. Most definitions in some way orient on the idea of getting AI to behave in a way that is 'good' rather than 'bad'. Thus framed, an obvious approach is to first specify a value function – a formal definition of what is good and bad – and then put AI to work towards optimizing for that value function. A value function might place weight on reducing human suffering, increasing wealth, decreasing inequality, and so on.

However, as soon as we try to specify what we value, it becomes clear that it is difficult or impossible to capture what we intend (Gabriel, 2020; Grossman and Hart, 1986; Hadfield-Menell and Hadfield, 2019; Krakovna et al., 2020; Russell, 2019; Wiener, 1960; Zhuang and Hadfield-Menell, 2020). Bostrom proposes some compelling thought experiments to illustrate this problem (Bostrom, 2014). Imagine our value function places weight on finding a cure for cancer. A super-powerful AI faithfully trying to optimize for our stated wishes could 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 might be most efficiently achieved by imprisoning humans and directly stimulating the neurons that trigger the experience of wellbeing.

These thought experiments are not isolated examples. It is in general difficult or impossible to write down a suitable value function because each concept, model or formalism is only an imperfect proxy (Aristotle, 2019; Heidegger, 1998; Kant, 1781; Kuhn, 1970; Plato, 2002; Wittgenstein, 1922). When we use concepts or formalisms to try to specify values, we are inevitably missing things. This results in the core problem identified in Section 1 of this paper: if AI does too good a job of organizing the world around a particular framing of values, then the richness of the world is homogenized to that framing.

With this perspective, we can turn the tables on the alignment problem. Instead of figuring out exactly what we value and trying to do a really good job of optimizing for that, the problem is to place boundaries on optimization for any particular conceptualization.

7.2 Other approaches

Before we go into more detail about our proposal, let's stop to consider some of the other solutions that have been suggested.

7.2.1 Inverse methods

The problem is that any particular way we have of specifying values is inadequate. One natural solution is to try to dynamically learn values from humans. This is a broad family called 'inverse methods', because

The problem is, still, if AI becomes too influential, with *whatever* particular form it currently has, the results are problematic. Let's consider a few specific ways this could manifest.

Stuart Russell's inverse RL approach. Our idea is similar (believing in your own uncertainty is a form of boundary). But *any fixed implementation of inverse RL* will have its own form of collapse. Any particular version of it is not true alignment.

Even our preferences, however we express them, are still just a proxy. 1) They will never fully capture what's in us (cf Steve's point about the deep wisdom in life), and 2) They don't express what's beyond us. Finally, 1 and 2 are actually coterminous, because the deeper you go inside us, the more it's beyond us.

7.2.2 Principles-based methods

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.

Is this 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.

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

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

Stuart Russell proposes assuming uncertainty and continuing to learn about what the values are. 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?

Iason Gabriel proposes a principles approach. But however we formalize the principles, we'll run into the same problem: they are still formalizations.

- What we want can be a poor proxy for what is advantageous for our wellbeing (Kerr, 1975)
- It is difficult or impossible to trade off between the present and future. It is plausible that there will be orders of magnitude more humans in the future, who are not yet born, than there are alive at present. Should we sacrifice our welfare now to promote their happiness (MacAskill, 2022)?
- Different people value different things (Sorensen et al., 2024). How do we weight these against each other? There is not even universal agreement on which principles are most appropriate for aggregating the preferences of different people.
- Should the value function we give to AI promote the welfare of humans only, or should it include other life on earth, or even undiscovered sentient species in the universe?

Our values are deeply embodied and intuitive, and difficult to fully capture in language (Anwar et al., 2024; Zhi-Xuan et al., 2024).

7.3 Our proposal

Alignment is not picking the right values or principles, or even the right system for inferring values or principles. Alignment is an evolving set of boundaries forming selective, intelligent constraints. Boundaries don't tell the system exactly what to do, and any particular boundary or set of boundaries is not a final answer.

This will lead to AI that is more 'intelligent' in a deeper sense. Not like paperclip intelligence, but like the wisdom of cells.

A common formulation of the alignment problem highlights the difficulty of writing down a complete, final specification of what humans value. In practice, algorithms that optimize concrete proxies or reward signals often produce solutions that satisfy the proxy while violating the intended goal ("specification gaming", "reward hacking", or proxy failure) (Amodei et al., 2016; Krakovna et al., 2020). This failure mode is not merely an engineering nuisance: as agents grow more capable, narrow objective optimization can drive dramatic, discontinuous failures where previously acceptable behavior collapses into pathological strategies.

'Doing what is good' vs 'doing what is good for us' vs 'doing what we want'

The problem is unique because AI will surpass our ability to control it. Sort of like the Founding Fathers writing the Constitution with its self-modifying ability. To set this future system, which is way out of their control, in a good direction.

The central problem of alignment is that AI systems do what we tell them to do.

We conjecture that the basic problem of alignment is that any particular form is not a complete answer. It's not possible to formalize our values. Three different angles to look at this from. First, our values are always changing. Think about fish versus Neanderthals vs humans. Second, we can't fully capture our values because they stretch below language into subtle, contextual intuition which probably even involves our bodies, our communities and so on. Third, what we want is not good for us in a larger sense.

Essentially our argument is the intersection of (big world hypothesis / etc) with (specification gaming / proxy failure / etc)

If you write down on a piece of paper the way you would like the world to be, and hand that piece of paper to me, and I do a *really* good job of making the thing you wrote happen, then the outcome will not be good for you. This is true no matter how cleverly you try to write down exactly what you want (Bostrom, 2014; Krakovna et al., 2020; Russell, 2019; Wiener, 1960).

Walk through the example of optimizing for likes, for reasoned likes, for wellbeing, for long-term wellbeing, for principles.

Any particular formalization is not the final answer.

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."

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

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

This has always been true, but it comes to a head with AI alignment because the optimizing force is so powerful. If we write down what we want, we'll get a paperclip outcome on some level. The crucial thing is that we make something more like an open-ended living system, and less like a paperclip-maker.

The point of alignment is not to say that any particular perspective is absolutely wrong or right.

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.

[Somewhere, walk through the idea of optimizing for thumbs up, then deliberative thumbs up, then wellbeing, then long-term wellbeing. see how there's still a problem even when we optimize for long-term wellbeing, defined in any particular way.]

Like with natural selection, gradient descent is a super powerful optimizing force that tends to collapse into myopic forms. This is a really key thing about AI. That in any kind of system, there's always a possibility for one form to take over (even a crystal is like that). But strong optimizers have this movement toward sucking more and more of the world into their formalism. Like fire.

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[1].

Giving ourselves what we want; the superorganism; increase correlations; the fermi paradox

The alignment problem is sometimes divided into two parts. The first part is specifying what we want. This is the 'specification problem' or 'outer alignment'. The second part is getting the AI to do what we've specified. This is 'inner alignment'.

(wait.. do i actually accept the inner vs outer classification? because my proposed method never actually tries to specify what is good...)

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).

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.

7.4 Objections

Q: Surely we want to make AI at least *somewhat* aligned to *human* values. 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: Sure. We don't deny the importance of all these local perspectives. It makes total sense that humans would want to advantage our own welfare, and we don't have a problem with that.

Q: Is this pure relativism? Everything is equal, you can't tell anything apart?

A: No. In fact, an important boundary is against excess relativism. 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. Boundaries protect against a particular perspective being taken to a dominating absolute.

7.5 Lifelike alignment

Boundaries mean much of our world is internal. Like replay in the brain: much of what's happening is "offline" - internal dynamics are only loosely perturbed by inputs. Even invertebrates are mostly internal dynamics.

[1] Cast another way, we could equivalently say that the aim of alignment *is* adhering to human values, but true human values cannot be fully be captured with formalisms.

Deacon has the idea that constraints reciprocally limit processes and give rise to life and consciousness (Deacon, 2012).

Indeed, the concept ‘concepts are incomplete’ is incomplete and will continue to evolve (Hofstadter, 1979).

7.6 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.

7.7 Awareness

From a subjective point of view, contextualization is awareness or "aboutness" (Yontef, 1993). Likewise, 'owning experience' or 'holding' is contextualization.

New boundaries are always needed as the environment changes. It fundamentally rests on awareness. Being aware of when any particular thing gets too concentrated, and placing a boundary to protect the "letting it be" unfolding.

By construction it's a deep mystery how contextualization works. If we have a particular, absolutely fixed strategy for regulating an obsession, then this strategy itself is part of the obsession. If we have an absolutely concrete idea about how alignment works, then that idea is part of the problem (importantly, it's also part of the solution; semi-permeable boundaries!).

Spiritual traditions suggest that the only 'absolute' truth is the self-evident truth of immediate experience: this could be viewed as not a truth in the normal sense, but something more like an orientation toward holding each perspective in awareness. That orientation itself is a system of boundaries. Allowing the potential that's inside us to creatively emerge. "The self who's in control is not any particular self but the unformalizable process itself."

7.8 Emergence of new form

And, new structure continues to emerge. This is true in the outside world, like with the arrival of replicators or of cells or of nervous systems. It's also true of our concepts and understanding. 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.

Sensitivity as a delicate balance between a lot of concepts. How could Beethoven write music? Allowing something greater than the self to operate, by being at the knife edge and not collapsing into one interpretation.

7.9 Beyond human level

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

Super-fish intelligence.

7.10 So what should we do?

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 life-life 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 life-like property of internal dynamics that applies contextualization/awareness to itself as the ultimate scalable boundary.

8 Acknowledgements

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

9 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. Barthes. *S/Z*. Éditions du Seuil, Paris, 1970.

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.

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. *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.

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.

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.

T. W. Deacon. *Incomplete Nature: How Mind Emerged from Matter*. W. W. Norton & Company, New York, 2012. ISBN 978–0393049916.

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.

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.

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.

U. Eco. *Opera aperta*. Bompiani, 1962.

W. Empson. *Seven Types of Ambiguity*. Chatto & Windus, London, 1930.

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.

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.

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.

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.

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

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

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. 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.

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.

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.

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

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

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.

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. The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science*, 20(2):231–237, 2009.

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

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.

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.

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.

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

A. Karnam, H. D. Kulkarni, N. R. Mutalik, and U. Gajendran. Obsessive-compulsive disorder with bizarre content of obsession. *Telangana Journal of Psychiatry*, pages 10–4103, 2025.

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.

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, 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.

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.

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.

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.

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.

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.

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.

H. Rauhut and J. Lorenz. The wisdom of crowds in one mind: How individuals can simulate the knowledge of diverse societies to reach better decisions. *Journal of mathematical Psychology*, 55(2):191–197, 2011.

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.

E. T. Rolls. Attractor networks in the brain. *Frontiers in Computational Neuroscience*, 16: 936306, 2022.

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/`.

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.

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

F. Schiller. *Über die ästhetische Erziehung des Menschen in einer Reihe von Briefen*. Verlag der Cotta'schen Buchhandlung, 1795.

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. 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, 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.

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.

J. R. Stroop. Is the judgment of the group better than that of the average member of the group? *Journal of experimental Psychology*, 15(5):550, 1932.

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.

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

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.

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

A. Traulsen and M. A. Nowak. Evolution of cooperation by multilevel selection. *Proceedings of the National Academy of Sciences*, 103(29):10952–10955, 2006.

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.

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.

E. Vul and H. Pashler. Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, 19(7):645–647, 2008.

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.

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

S. A. West, A. S. Griffin, A. Gardner, and S. P. Diggle. Social evolution in microorganisms. *Nature Reviews Microbiology*, 4(8):597–607, 2006.

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.

D. S. Wilson. *The Natural Selection of Populations and Communities*. Benjamin/Cummings Pub. Co., Menlo Park, Calif., 1980. ISBN 978–0805393603.

W. K. Wimsatt and M. C. Beardsley. The intentional fallacy. *The Sewanee Review*, 54(3): 468–488, 1946.

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

C. Wood. Alien calculus could save particle physics from infinities. *Quanta Magazine*, April 2023. URL `https://www.quantamagazine.org/alien-calculus-could-save-particle-physics-from-infinities-20230406/`.

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.

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

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.

K. Zhang. Understanding directional selectivity in rat head-direction cells in a theoretical framework. *Journal of Neuroscience*, 16(6):2112–2126, 1996.

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.