

# Wireheading and the Conditions of a Common World

Attention, Reward, and the Limits of Universal Self-Promotion

Flyxion

Independent Researcher

October 2026

## Abstract

Wireheading, in its original sense, is the direct activation of a reward mechanism in place of the activities through which that mechanism ordinarily operates. This essay argues that the concept names a structural hazard that extends well beyond pleasurable stimulation. The hazard arises when a signal of success becomes independently optimizable while the conditions that gave the signal its significance are left to deteriorate. Applied to contemporary institutions, the claim requires care, because pleasure, entertainment, publicity and digital communication are not intrinsically pathological. The relevant question is when the production of rewarding signals begins to displace the activities and relationships those signals were meant to serve. The essay develops a small formal model of the allocation of effort between substantive work and signal production, examines the thought experiment of a society in which everyone follows the advice to become a broadcaster, and draws attention to the maintenance labour on which every visible activity depends. It considers institutional cases and then hypernormalization, the process by which a drifted measure is stabilized through the universal reproduction of its form, which obstructs correction either by concealing private doubt or, where dysfunction is openly acknowledged, by leaving knowledge without a means of organized change. It states a material limit on the labour that can be allocated outside provisioning, and two requirements of viability, one concerning the availability of dissent, illustrated by the premise of *Minority Report*, and one concerning the diversity of modes of work. It then distinguishes the documentation of an activity, which can add to its value, from the organization of an activity around its publishability, which can distort it, and illustrates the distinction with two cases: recordings of ordinary domestic labour as demonstration data for robotic manipulation, and a class of lightweight, paper-based, string-actuated robots designed to make their own construction and upkeep intelligible. After addressing the objection from paternalism, the essay proposes a universalizability test for attention strategies and a set of design questions concerned with preserving the possibility of

correction. Its conclusion is that the deepest danger of wireheading is the progressive loss of the capacity to notice and repair the difference between feeling rewarded and living well.

## 1. From Reward to Its Artificial Satisfaction

The experiments of Olds and Milner in the early 1950s established that rats would press a lever in order to deliver electrical stimulation to certain regions of their own brains [22]. The result was striking for what it suggested about the architecture of motivation. Reward, in an ordinary organism, is a regulatory signal that is activated through engagement with a world: through eating, exploring, mating, solving, caring. The electrode short-circuits the engagement and delivers the signal directly. The signal remains, but the conditions that made it informative have been removed. In the vocabulary that later became standard, the organism has been wireheaded.

The term has since travelled. In the literature on reinforcement learning and artificial agents, it describes a system that modifies its own reward channel, or the sensory inputs from which reward is computed, rather than improving its performance at the task the reward was designed to elicit [24, 9]. Related phenomena go under the names of reward hacking and specification gaming [1, 32]. In moral philosophy, Nozick’s experience machine poses a cognate question about whether the felt quality of experience is the only thing that matters, and his negative verdict, that most people would decline to plug in, is usually taken to show that we care about actually doing and being certain things and not merely about the experience of doing and being them [21]. Neuroscientific work distinguishing the wanting of a reward from the liking of it adds a further complication: the machinery that drives pursuit is dissociable from the machinery that registers pleasure, so that a system can be strongly motivated toward a signal that no longer delivers what it once did [3].

This essay is concerned with a broader and more deliberately analogical use of the concept. A person watching videos is not an organism with an electrode in its septal region, and an institution pursuing publicity is not an agent modifying its own reward function in the technical sense. The analogy concerns the *separation of reinforcement from the conditions that would make reinforcement informative*. Where such a separation opens, the system acquires the ability to register success without acquiring a corresponding ability to sustain whatever success was supposed to mean. The governing question of the essay is accordingly the following. What happens when a system becomes increasingly effective at producing and detecting signals of success, while its effectiveness at maintaining the activities, relationships and material conditions to which those signals referred does not keep pace, or declines?

Two cautions frame the discussion. First, the analogy is not offered as a moralistic indictment of pleasure, leisure, entertainment or publicity. Each of these can be, and very

often is, part of a worthwhile life and a well-functioning society. The argument concerns a structural relationship between signal and activity, and it can be stated without any assumption that screens are bad or that everyone ought to be perpetually productive. Second, the analogy has limits. Where the essay describes institutions as wireheaded, it means that their reinforcement structures exhibit a particular gap, and it does not mean that their members are deluded, cynical or complicit. Much of the interest of the topic lies in the fact that the gap can open and widen among participants who are sincere and diligent.

## 2. The Signal and the Activity

Recognition can indicate worthwhile achievement. Revenue can help sustain useful production. Audience engagement can accompany meaningful communication. In each case the signal is correlated with an underlying good, and the correlation is the reason the signal is worth pursuing and worth using as a basis for allocating resources. The difficulty is that each signal also admits strategies that improve its measurement without improving its object. This regularity has been stated in several forms: Goodhart’s observation that a statistical regularity tends to collapse once it is used for control, Campbell’s law that the more a quantitative indicator is used for social decision making the more it distorts the processes it was intended to monitor, and Strathern’s compact formulation that when a measure becomes a target it ceases to be a good measure [13, 5, 35]. Muller’s book-length treatment documents the pattern across education, medicine, policing and business [20].

It is important to see why the problem is stronger than dishonesty. A common response to metric distortion is to suppose that it results from gaming by unscrupulous actors, and that it could be eliminated by better monitoring and sterner penalties. The more troubling cases are those in which participants sincerely follow the incentives available to them. A researcher who writes proposals in the style that funding panels reward, a teacher who prepares students for the form of an examination, and a creator who learns which thumbnails are clicked are all responding rationally to the signal as it is presented. The distortion emerges at the level of the system when the incentives systematically reward the production of appearances, responses or measurements more reliably than the work they are meant to recognize.

A minimal model makes the structure explicit. Consider an agent with a unit budget of effort, to be divided between substantive work  $w$  and signal production  $p$ , so that  $w + p = 1$  with  $w, p \geq 0$ . The substantive value produced is  $V = aw$  for some  $a > 0$ . The observable signal is  $S = bw + cp$ , where  $b > 0$  measures how well substantive work is registered by the signalling channel and  $c > 0$  measures how effectively direct signal production is registered. The agent is rewarded by the signal, and bears a convex allocation cost  $\frac{\kappa}{2}(w^2 + p^2)$  with

$\kappa > 0$ , which represents the diminishing returns of concentrating all effort on a single kind of activity. The agent therefore chooses  $w$  to maximize

$$U(w) = bw + c(1 - w) - \frac{\kappa}{2}(w^2 + (1 - w)^2), \quad w \in [0, 1].$$

**Proposition 1.** *The optimal allocation is  $w^* = \text{clip}(\frac{1}{2} + \frac{b-c}{2\kappa}, 0, 1)$ . Consequently  $w^*$ , and with it the substantive value  $V^* = aw^*$ , is non-increasing in  $c$  and is non-decreasing in  $b$ . If  $c - b \geq \kappa$ , then  $w^* = 0$  and no substantive value is produced.*

*Proof.* The function  $U$  is strictly concave in  $w$ , since  $U''(w) = -2\kappa < 0$ . Setting  $U'(w) = b - c - \kappa(w - (1 - w)) = b - c - \kappa(2w - 1) = 0$  gives the interior solution  $w = \frac{1}{2} + \frac{b-c}{2\kappa}$ . Strict concavity implies that when this value lies outside  $[0, 1]$  the maximizer is the nearest endpoint, which yields the clipped form. Monotonicity in  $b$  and  $c$  follows directly, and the threshold condition follows from setting  $\frac{1}{2} + \frac{b-c}{2\kappa} \leq 0$ .  $\square$

The proposition is elementary, but the interpretation is not. The parameter  $c$  is not fixed by nature, and it is worth considering how it might vary with the competitive environment. One particular escalation scenario arises where the signal is positional, so that rewards depend on rank and not on level. If the effectiveness of promotional effort rises with the number of rivals who also promote, for instance because each competitor must match the promotion of the others merely to hold its place, then one may write  $c = c_0 + \gamma(n - 1)$  for  $n$  competitors and some  $\gamma > 0$ . This form is an additional assumption and is not derived from competition as such. Crowding can equally reduce the effectiveness of promotion, since audiences saturate and distinguishing marks lose their force, and in some environments  $c$  would fall with  $n$ . Under the escalation assumption, the proposition implies that substantive effort declines as the field becomes more crowded, even though no individual has become less honest or less capable. Hirsch's analysis of positional goods and Frank and Cook's account of winner-take-all markets describe dynamics of this kind in other vocabularies [16, 12]. Rosen's economics of superstars supplies a related explanation of why small differences in measured standing can translate into large differences in reward, which can raise the returns to marginal signal production in some settings [25].

The model also clarifies what is and is not being claimed. It does not claim that  $c > b$  in every domain, nor that signal production is worthless. In domains where the signalling channel registers substantive work well, so that  $b$  is large relative to  $c$ , the system rewards substance and behaves well. The claim is that the relationship between  $b$  and  $c$  is a contingent and revisable property of the institutional environment, that it can drift, and that participants following local incentives will not in general be positioned to notice the drift.

### 3. Attention as Both a Resource and an Obligation

A prominent piece of contemporary advice holds that everyone should cultivate an audience, build a personal brand and maintain a continuous public presence. The advice has an evident basis. Neglected work does sometimes remain neglected for want of any means of being found, and publicity can correct that. Public communication also helps to coordinate people who would otherwise not know of one another's existence or interests. These benefits are real, and an account of attention that ignored them would be tendentious.

Nevertheless, attention has features that complicate its treatment as an ordinary resource to be acquired. Herbert Simon observed in 1971 that a wealth of information creates a poverty of attention, and that the design problem of an information-rich world is the efficient allocation of a scarce capacity among the sources that might consume it [31]. Attention is finite per person, and it cannot be accumulated or stored. It is also unusually rival in a particular sense: attention directed to one source is attention withdrawn from every other. A strategy that succeeds for one participant by capturing more attention therefore succeeds, to the extent that the total is fixed, by reducing what is available to others. Wu's history of the commercial capture of attention describes how successive media industries have exploited exactly this structure [37], and Debord's older analysis of the spectacle describes the condition in which lived relations are replaced by their representation [7].

This suggests a distinction between two kinds of communicative effort. Some communication helps people coordinate: it conveys what a person can contribute, what a community needs, how a technique is performed. Other effort chiefly changes who gets noticed, and has little effect on what there is to notice. The first kind has positive-sum possibilities, since a successfully coordinated pair of parties are both better off. The second kind has the character of a contest whose aggregate return, to the extent that attention is fixed, is bounded by the total available, while its aggregate cost is the sum of everyone's expenditure. Most actual communication mixes the two, and the essay does not claim that they can be cleanly separated. The claim is that the proportion matters, and that a social environment can raise the share of effort devoted to the second kind without anyone intending it.

A related observation concerns the relation between contribution and recognition. Curtis contrasts the committed, largely anonymous participation he attributes to the civil rights movement with a contemporary expectation that a commitment be made publicly visible, since visibility is now a main source of validation [4]. His historical comparison is sweeping and would need qualification, since the movement had prominent figures as well as anonymous participants. The conceptual point is nevertheless pertinent: the wish to contribute can become entangled with the requirement to be recognized as contributing, so that the second begins to shape what is done in the name of the first.

It follows that an individual's successful publicity strategy cannot automatically be generalized into a prescription for everyone. The fact that a given person has benefited from being visible is evidence that visibility can be beneficial for that person under those circumstances. It is not evidence that universal visibility would be beneficial, or even possible. The next section makes this precise.

#### 4. The Universal Broadcaster

Imagine a society in which everyone follows the advice to become a broadcaster. It would be a mistake to assume that catastrophe follows simply because people make videos. There is nothing intrinsically ruinous about recording and sharing, and a society in which a large fraction of the population communicated publicly might be richer in some respects than one in which few did. The thought experiment is useful for a narrower reason: it exposes the practical dependencies of broadcasting and the incompleteness of broadcasting as a universal solution to livelihood.

Broadcasting requires food, electricity, networks, buildings, equipment, maintenance and care. It also requires audiences whose attention is not already exhausted. Let a population of  $N$  persons each possess a time budget  $T$ . Suppose a fraction  $\sigma$  of total time must be devoted to sustaining labour, meaning the production of food, the operation and repair of infrastructure, and the care of the young, the ill and the old, and suppose a fraction  $\varphi$  of the population broadcasts, each broadcaster devoting a fraction  $\tau$  of their time to production. The time available for viewing is then at most  $(1 - \sigma - \varphi\tau)NT$ , and the mean viewing time received per broadcaster is

$$r(\varphi) = \frac{(1 - \sigma - \varphi\tau)T}{\varphi}.$$

**Proposition 2.** *Assume  $T > 0$ ,  $0 \leq \sigma < 1$  and  $\tau > 0$ , and let  $r(\varphi) = (1 - \sigma - \varphi\tau)T/\varphi$  be defined on the feasible range  $0 < \varphi \leq \min\{1, (1 - \sigma)/\tau\}$ . Then  $r$  is strictly decreasing in  $\varphi$ , and  $r(\varphi) \leq (1 - \sigma)T/\varphi$ . If in addition  $\sigma + \tau \leq 1$ , then  $\varphi = 1$  is feasible and  $r(1) = (1 - \sigma - \tau)T$ , which is smaller than  $r(\varphi)$  for every feasible  $\varphi < 1$ . The quantity  $r(\varphi)$  is an upper bound on mean viewing time per broadcaster, and equals that mean if and only if all time not devoted to sustaining labour or to production is spent viewing broadcasts.*

*Proof.* Differentiating,  $r'(\varphi) = -T(1 - \sigma)/\varphi^2 < 0$  for  $\sigma < 1$ , which establishes strict monotonicity. The bound follows from  $1 - \sigma - \varphi\tau \leq 1 - \sigma$ . Non-negativity of viewing time requires  $\varphi\tau \leq 1 - \sigma$ , which together with  $\varphi \leq 1$  gives the stated feasible range, and  $\varphi = 1$  lies in that range exactly when  $\tau \leq 1 - \sigma$ . At  $\varphi = 1$  the expression reduces to  $(1 - \sigma - \tau)T$ , and strict monotonicity gives the comparison with smaller feasible values of  $\varphi$ . The characterization of the upper bound follows because  $r$  was constructed from the total time not committed to sustaining labour or production.  $\square$

The proposition concerns means, and it establishes attention dilution under the stipulated time budget and no more. It does not show that universal participation in broadcasting is impossible, nor that audience-based livelihoods are unattainable. A conclusion about income would require further assumptions about how attention is converted into revenue and about the income each broadcaster needs in order to live, and the model makes neither. Attention is also in practice distributed very unequally across broadcasters. In heavy-tailed environments a small number of participants receive a large share of the attention, so that the median experience lies below the mean [25, 12], and a mean figure may therefore overstate what a typical participant receives. The model’s purpose is the narrower structural point that any livelihood derived from audience attention depends on audiences whose time is not wholly committed to broadcasting, and on a population of workers, not themselves full-time broadcasters, who sustain the conditions of both. A strategy can depend on others not adopting it to the same degree.

A related distinction applies to a familiar promise, namely that anyone can escape ordinary employment by becoming a coach, an influencer, or a seller of advice about escaping ordinary employment. The word “anyone” is ambiguous. An opportunity can be open to any individual without being simultaneously available to everyone, just as any individual may seek a seat in a theatre although not everyone can occupy the front row. The strongest target of the present argument is therefore advice that presents audience-based income as a general solution while leaving unexplained its dependence on differentiated labour and on audiences supplied by others. Advice of this kind often has a recursive structure in which the product sold is the strategy of selling the product. Where the revenue of the advisers derives largely from those seeking to become advisers, the arrangement requires a continuing inflow of new entrants and cannot be extended to the whole population without changing its character. This is not to say that any particular adviser is insincere. It is to say that the form of the offer, taken as a general route to freedom, depends on a distribution of labour and attention that its own universal adoption would alter.

## 5. The Hidden Work of Sustaining Experience

The preceding section turns attention from the visible content of a culture to the conditions that make its production and consumption possible. A video must be recorded on a device that was manufactured, transmitted over a network that is maintained, stored in a facility that is powered, and watched by a person who has eaten, slept, and been cared for. The activities that supply these conditions include infrastructure repair, sanitation, food production, caregiving, teaching, and ecological maintenance. A distinctive feature of such activities is that their benefits are often less conspicuous than their interruption. A

functioning water system attracts no attention until it fails.

This asymmetry of visibility has been recognized in several traditions. Mierle Laderman Ukeles's 1969 manifesto of maintenance art distinguished a development system, associated with individual creation, novelty and change, from a maintenance system, which keeps the results clean, preserved and sustained, and it pressed the question of who attends to the everyday labour that makes creative work possible [36]. Susan Leigh Star's work on the ethnography of infrastructure observed that infrastructure is by its nature invisible until it breaks, and that this invisibility shapes how the labour involved is valued [34]. More recent historical scholarship on maintenance and repair has argued that innovation receives disproportionate attention relative to the upkeep on which most actual technological life depends [27]. James Scott's account of legibility makes a related point from the side of administration: what is easily measured and displayed tends to be favoured over what is locally embedded and difficult to summarize [29].

A distinction drawn in an interview with the scholar Andrew Russell sharpens the asymmetry. Russell separates innovation as an activity, meaning invention and creation, from talk about innovation, and his criticism concerns the attention that the language attracts, since time given to it competes with time for maintenance [28]. The point fits the structure of this essay. Innovation can register worthwhile change, but it can also become a designation that institutions pursue independently of whether their changes improve anything, while effective maintenance attracts little recognition precisely because it prevents conspicuous events. The criticism is not a blanket preference for preserving existing technologies, and it does not oppose technological ambition. It also qualifies any simple valorization of upkeep. Preserving essential capacities, such as sanitation, scientific instruments or shared knowledge, can enable substantial change, whereas maintaining an institution's appearance of competence, or an existing distribution of power, can obstruct it. The relevant questions are what is being maintained, for whom, and with what capacity for revision.

The argument does not divide society into virtuous manual workers and frivolous cultural workers. Art, research, entertainment and conversation can themselves sustain a worthwhile common life, and some of the most important maintenance in a society is cultural and intellectual: the preservation of languages, the transmission of skills, the upkeep of shared standards of evidence. A society that left all of this unmaintained would be impoverished in ways that no amount of material provision would repair. The distinction that matters is therefore not between kinds of occupation. It is between activities that contribute to a common life and activities that increasingly consume the supporting conditions of that life merely in order to reproduce engagement. A painter whose work enriches a community is on the right side of this distinction, and so is a repairer of pipes. A publicity apparatus

whose output is chiefly more publicity, and which draws labour and attention away from the maintenance of its own supports, is on the other side.

Graeber’s discussion of occupations whose holders themselves doubt their usefulness is relevant here as an example of how the relationship between reward and contribution can come apart at the level of individual work [15]. The present essay is more cautious than Graeber’s in one respect. Self-assessments of usefulness are an unreliable guide, and the difficulty to be diagnosed is structural and not psychological. The question is whether the reward structure of a field tracks the contribution of the field, and this can be examined without relying on the opinions of those who work in it.

## 6. Documentation and Display

The wireheading argument is easily misapplied to recording as such, and a clear distinction is needed. The concern raised by a society in which people constantly broadcast their most publishable moments as photographs and videos is not that recording is bad. It is that a particular selection process operates on what is recorded and shared. Repetition, preparation, failed attempts and maintenance tend to disappear from public view, although they constitute much of what makes any accomplishment possible. People then encounter a public world composed disproportionately of exceptional moments, and some begin arranging their own lives in order to supply such moments. The argument becomes more exact when it concerns this feedback loop between display and the organization of activity, and less exact when it concerns screen time or recording in general.

The distinction can be stated as a question about the direction of dependence. Either the feedback from recording improves the capacity to act, or action is increasingly organized around obtaining the feedback. In the first case documentation is subordinate to the activity: it serves as a diagnostic, as a means of instruction, as a way of inviting correction. In the second case the activity is subordinate to its documentation, and the form of the activity is progressively shaped by what will display well.

### 6.1 *Recorded labour as demonstration*

A concrete illustration is provided by ordinary domestic and municipal labour. A recording of someone collecting litter along a roadside, or washing a stack of dishes, is not an optimized display of exceptional moments. It captures repetition, partial failures, regrasping, the handling of awkward objects and the sequencing of tasks. Such material can be useful in several distinct ways. It can communicate practical technique to a human learner. It can reveal labour that is otherwise overlooked, and thereby alter how that labour is regarded. And it can provide examples for the study of robotic manipulation, a field in which tasks of

this sort remain difficult. The literature on learning from demonstration treats recorded human performance as a primary source of training signal precisely because the tasks are hard to specify analytically [2].

The usefulness of such recordings depends on what they capture and on how someone can use them, and the claim made here is a modest one. Raw video is not training data in any straightforward sense, and the translation from observed human movement to robot control involves substantial difficulties of embodiment and correspondence. The relevant claim is limited. Broadcasting is not inherently a withdrawal from useful work, and a recording that documents mundane labour in an unoptimized way sits near the opposite end of the spectrum from the curated highlight. Its value is largely independent of how many people view it, and does not depend on the performer having organized the work around the camera.

## ***6.2 Lightweight robotics and intelligible construction***

A second illustration concerns the design of machines. Consider a class of lightweight robots in which the body is constructed principally from folded or laminated paper, internal strings transmit pulling forces to produce motion, and pneumatic components supply or regulate actuation. For convenience these may be called pneumatic endo-marionettes. The general family to which such designs belong includes tendon-driven mechanisms, origami-based fabrication and soft robotics [10, 26]. The precise arrangement of strings, bladders and valves would matter a great deal in any real design, and the present discussion does not depend on it. What is of interest is the connection between this research direction and the observation of ordinary tasks. Observing how a person washes a dish or lifts a bag invites the question of what movements, compliance and contact with the environment a mechanism would need in order to accomplish something useful, and a lightweight body with distributed internal tension is a natural candidate for exploring that question cheaply.

Here the distinction between documentation and display acquires a design interpretation. Filming a prototype can help diagnose its failures, teach someone how to build it, or invite a useful correction. The same prototype could also be optimized for an impressive clip while remaining unreliable outside the demonstration, tuned for the one successful take and brittle everywhere else. The first relationship between camera and machine builds capacity. The second builds the appearance of capacity. A research programme can be assessed in part by asking which of these relationships its working practices tend to produce.

The concept of a *didactic* robot sharpens the point. A machine might be designed so that its own construction and maintenance are made understandable through its behaviour. Suppose such a robot prepared itself by applying sheets of paper to its body one at a time, so that the manner in which its structure is assembled and repaired is displayed in the course of ordinary operation and not concealed beneath a finished exterior. The preparation would

then be part of what is worth attending to, and not a hidden precondition. Imagine further that a child is present while the robot, balanced on one leg, struggles to pull on a pair of trousers and comments humorously on its predicament. The child can see the relationship between balance, movement and the difficulty of dressing. The humour arises directly from what the robot is doing, and does not require a separate stream of entertainment laid over the activity. Limitations become occasions for shared attention and learning.

This case reverses the publishable-moment problem. Under the logic of display, preparation, repair and failure are edited out, because they do not present well. In the didactic design they are placed at the centre, because they are where the instructive content lies. The robot teaches through the care it takes with ordinary tasks, including the care required to keep itself working. The lesson generalizes. Entertainment, humour and visibility are compatible with capacity-building when they arise *from* the intelligible performance of an activity. The concern is with systematic displacement, in which the production of display comes to replace the activity and erodes the capacities that support it. It is not that every entertainment must itself develop capacities, a requirement the discussion of leisure below rightly rejects.

## 7. Institutional Wireheading

The analysis extends well beyond social media. Several institutional cases share the analytical structure of a widening gap between institutional reinforcement and institutional purpose. Each would require detailed evidence in a full empirical treatment, and the following descriptions are offered as schematic illustrations of a pattern whose presence in particular institutions must be established case by case.

A university can become highly effective at producing rankings, publicity and competitive grant proposals while reducing the time available for teaching and inquiry. Strathern's account of audit culture in British higher education describes how the demands of accountability, once internalized, reorganize the activity they were intended to oversee [35]. A business can improve customer engagement metrics while making its service harder to leave, so that the measured attachment of customers reflects the cost of departure and not the value of the service. Hirschman's analysis of exit, voice and loyalty is the classical statement of why the ability to leave matters as a mechanism by which organizations learn of their own decline [17], and an organization that suppresses exit forfeits the correction it would otherwise receive. A public agency can satisfy reporting requirements without resolving the difficulties those reports describe, so that the production of compliant documentation becomes the effective task.

The shared structure admits a general description. Let  $P$  denote the purpose of an

institution and  $M$  the measurement regime through which its performance is registered and rewarded. Initially,  $M$  is a reasonable proxy for  $P$ . Participants optimize against  $M$ , since that is what the institution rewards. Because  $M$  is a coarser object than  $P$ , there are directions in which  $M$  can be improved at little cost to  $P$  or even at the expense of  $P$ . As optimization proceeds, the relationship between  $M$  and  $P$  changes: the proxy relationship, which held under the earlier distribution of behaviour, ceases to hold under the new one. The institution continues to register improvement, and the registration appears to confirm that the institution is succeeding. Nothing in the reinforcement signal itself indicates that the correlation has failed.

It is worth emphasizing that this process requires neither coordinated deception nor universally cynical participants. Each participant responds to the signals they face, and the system as a whole drifts. The point has an uncomfortable implication: moral exhortation directed at individual participants will have limited effect, because the relevant error is located in the structure of the reinforcement. This is also why the diagnosis has a positive counterpart. If the gap is structural, it can be addressed by structural means, and these are discussed in a later section.

## 8. Hypernormalization and the Closure of Correction

The gap between signal and purpose described in the preceding section becomes more difficult to repair when it is stabilized by a further process, for which the term *hypernormalization* is available. The term derives from Alexei Yurchak’s anthropological study of the last Soviet generation, where it names a specific development in late socialist public life [38]. Yurchak observed that after the death of Stalin the authoritative discourse of the state, in its slogans, meetings, reports and rituals, became increasingly standardized and increasingly reproduced in a fixed form. Because the form was now fixed, participants could, and did, reproduce it exactly without any need to treat its literal content as a claim to be believed or disbelieved. Yurchak describes this as a shift from the *constative* to the *performative* dimension of ritualized speech: attending a meeting, voting in the approved manner or repeating an approved phrase became acts that were valuable as acts of participation, and their propositional meaning receded. The result was neither mass belief nor mass cynicism. It was a condition in which the form of public life was reproduced with great fidelity while the relationship between that form and anyone’s actual assessment of the world became largely irrelevant to the reproduction.

Yurchak is careful to resist two simplifications that popular uses of the term have encouraged. The first is the binary of sincere believers and secret dissidents, a picture on which people either endorsed the system or privately opposed it. His material shows that

many participants found real meaning, friendship and creative life within and alongside the reproduced forms, and that the performative shift opened spaces of activity that the constative reading of the discourse would have prohibited. The second simplification is the assumption that the system’s stability was straightforwardly the result of fear. A good deal of the stability came from the fact that the reproduced forms were the only publicly legible ones, so that every participant had reason to expect every other participant to continue reproducing them. The term was later extended to Western societies in Adam Curtis’s 2016 documentary *HyperNormalisation*, which uses it to describe a condition in which widespread awareness that institutions are failing coexists with the absence of any credible alternative, so that a simplified picture of the world is maintained because no other framework seems available [6]. That extension is looser than Yurchak’s, and is Curtis’s own interpretation. It should be treated as a suggestive use of the term and not as a finding about contemporary societies, and it should not be read as a direct account of Yurchak’s book. In a podcast interview about the film, Curtis glosses Yurchak’s observation as the condition in which everyone knows that things are not right, the leaders know that everyone knows, and nothing is done because no other picture of the world is available. He adds that he is not claiming that contemporary societies resemble the Soviet Union, only that they share a sense of unreality and of leaders who lack control [4]. This is again his paraphrase and not Yurchak’s own formulation. The present essay adopts only the structural core of Yurchak’s account, and applies it with attention to the differences between the cases.

The structural core can be connected directly to the model of the preceding sections. Suppose that a measurement regime  $M$  has drifted away from its purpose  $P$ , and suppose that many participants have private knowledge of this, whether in the form of a settled judgment or of an inarticulate unease. Whether that knowledge becomes public depends on the cost of expressing it relative to the cost of continuing to reproduce the form. Where expression is costly, for instance because recognition, employment or standing depend on visible compliance with  $M$ , each participant has reason to reproduce the form regardless of their private assessment. Kuran’s analysis of preference falsification describes the resulting equilibrium: public preferences diverge from private ones, and because everyone observes only public behaviour, each participant underestimates how many others share their doubts [19]. The effect is a form of pluralistic ignorance sustained by the very behaviour that conceals it.

*Remark 1.* Let  $g_i$  denote participant  $i$ ’s private assessment of the gap between  $M$  and  $P$ , and let  $B_i$  denote  $i$ ’s publicly observable behaviour with respect to  $M$ . Suppose each participant faces a cost  $d_i$  of public deviation and a benefit  $e_i$  from expressing  $g_i$ , and reproduces the standard form whenever  $d_i > e_i$ . If  $d_i > e_i$  holds for every participant, then  $B_i$  takes the same value for all  $i$  whatever the values of  $g_i$ , so that observed behaviour

carries no information, through that channel, about the distribution of private assessments. Other evidence, including gradual changes in outcomes, may of course remain available. A system whose members are all in this regime exhibits maximal observable agreement together with possibly large hidden divergence. If the inequality fails for some participants, their public dissent may lower the cost of deviation for others. That is a plausible additional mechanism, of the threshold kind familiar from models of collective behaviour [14], and it is not established by the remark itself.

The remark helps to explain two features that Yurchak and later commentators have emphasized. The first is that apparent stability is poor evidence of health. A regime in which every observable indicator is favourable may be one in which the indicators have ceased to be informative, because they are produced by a process that no longer depends on the state of the thing indicated. The second is the possibility of abrupt change. Yurchak's title records the experience of the late Soviet period, in which the order seemed permanent until it ended, and many observers on all sides were surprised by the manner of the end. A system that has lost its internal channels for registering divergence may conceal accumulating divergence until a disruption makes it publicly visible.

Two obstructions to correction are distinct. The first is concealed disagreement, the condition modelled in the remark, in which public compliance hides the distribution of private judgments. The second is the condition emphasized in Curtis's account, in which dysfunction is already widely acknowledged, and is known to be acknowledged, but the acknowledgement does not issue in action because no credible alternative or capacity for collective action is available. In that condition exposure loses its disruptive force, since each further revelation confirms what people already expect and does not change what they think possible. Protected dissent addresses the first obstruction and does little for the second, which also requires institutions, coordination and sufficiently developed proposals through which people can do something with what they know. In the interview he points to the large London march against the Iraq war in 2003, whose participants were sincere and numerous but who, in his telling, then went home, and he contrasts this with sustained commitment of the kind he attributes to the American civil rights movement. He further suggests that Occupy and the Arab Spring stalled because they confused process with content, that is, because they mistook their procedures of assembly and participation for an answer to the question of what should replace the arrangements they opposed [4]. These are his interpretations of particular episodes and are not established here, but they supply a useful test for any activity: whether it merely expresses dissatisfaction, or develops the means to change the conditions that produce it.

The notion therefore describes a second-order form of the wireheading structure. In the first-order case, a signal of success is optimized in place of the activity it indicated. In the

hypernormalized case, the signal that is being optimized is the appearance of normality itself, the continued smooth reproduction of expected forms. Participants are rewarded for not disturbing it, and the reward is real, since the stability of the form is a precondition of their own standing. In the concealed-disagreement case, correction is then doubly obstructed. The first-order gap between  $M$  and  $P$  is concealed by the reproduction of  $M$ , and the second-order evidence that the gap exists, namely the private doubts of participants, is concealed by the reproduction of the form of agreement. In the acknowledged-dysfunction case the evidence is not concealed, and the second obstruction is of a different kind: available knowledge is not translated into organized change.

Contemporary communication environments exhibit a version of this structure at the level of ordinary social presentation. A feed composed of publishable moments establishes a standard of normal life that few participants actually inhabit, and because each participant contributes to the standard while privately knowing the gap between it and their own experience, each is positioned to believe that the others live up to it. The mechanism is the same as in Kuran's analysis, applied to the presentation of lives and not to the endorsement of policies. A different mechanism is emphasized in Curtis's interview, where he argues that systems which return to users more of what they are already known to like are conservative, since they hold the informational situation of each group stable and conceal from each what the others know [4]. This is an interpretive claim about recommender systems and not an established result, but it indicates a further route by which a signal of engagement can stabilize the arrangement that produces it. The comparison should not be pressed too far. Contemporary media environments lack the central authority of the Soviet case, the cost of deviation is typically social and not coercive, and the forms reproduced are decentralized products of many competing sources and not of a single discourse. The comparison nevertheless indicates what to look for: a regime in which every visible indicator is favourable and in which the capacity of the system to register the gap between its indicators and its purposes has quietly disappeared.

The point also bears on the positive proposals to be developed below. If hypernormalization consists in the conversion of private knowledge of a gap into universal public reproduction of a form, then a corrective must lower the cost of making private knowledge public, or must supply channels through which it can inform decisions without requiring any individual to bear the cost of open deviation. This is a different requirement from the introduction of better metrics, since a new metric will itself be reproduced in the same manner if the conditions of reproduction are unchanged.

## 9. Material Limits and Corrective Requirements

A further question concerns what a society can carry. If wireheading in the extended sense is a condition in which reward has separated from the sustaining of what reward referred to, one may ask how much of a society's activity can be of that kind. Three considerations bear on the question, and they have different evidential standing. One is a material limit that follows arithmetically from a stated provisioning model. The other two are requirements of viability, conditions that a self-correcting society must meet, for which the analysis supplies no quantitative thresholds. The discussion concerns functions and allocations of labour and not classes of persons, since the same individual may occupy different functions at different times.

The material limit follows from the dependence of every visible activity on sustaining labour, discussed earlier. Let  $N$  be the population and  $L \leq N$  the labour supply, measured in units of full-time labour over a period. Let  $L_s$  be the labour allocated to provision, meaning the production of food, the operation and repair of infrastructure, and the care required to sustain the population, and let  $\lambda$  be the number of persons whose needs are met per unit of provisioning labour, in the same units. Provision for the whole population requires  $\lambda L_s \geq N$ , so that  $L_s \geq N/\lambda$ . The labour remaining for everything outside the provisioning sector is therefore at most

$$L - \frac{N}{\lambda}, \quad \text{or as a fraction of the labour supply,} \quad 1 - \frac{N}{\lambda L},$$

which equals  $1 - 1/\lambda$  under the normalization  $L = N$ , that is, one unit of labour per person, and is smaller when  $L < N$ , as when some persons are dependents. The expressions assume  $L > 0$  and are meaningful only if  $\lambda L \geq N$ . If  $\lambda L < N$  they are negative and denote a provisioning shortfall and not an available residual. Because the accounting is in units of labour, mixed roles and part-time work are accommodated. The residual measures the capacity to support activities outside provisioning in general: education, research, art, play, governance, and care not counted as provision, together with any activity whose rewards have become detached from its purposes. It must not be read as a quantity of wireheading, since most of what it supports may be valuable. The limit says only that the capacity to support activity beyond provisioning is finite and depends on the productivity of provisioning labour. That productivity,  $\lambda$ , is not fixed. It varies with technology, and advances in automation may raise it, which is one reason the design of machines that perform and display ordinary maintenance bears on the question.

The first requirement of viability concerns the capacity to detect drift between signal and purpose. This capacity cannot be derived from material accounting. It requires that some participants are placed to notice divergence, that they are independent of the reward

structure being assessed, that they have access to the relevant information, and that they have the authority or connections through which their observations bear on decisions. These properties are not captured by a numerical proportion of the population, since a few well-placed participants may suffice where many without standing do not. The requirement is therefore qualitative, a matter of authority, independence, access and connection, and the analysis supplies no proportion for it. Its failure can occur in a society that is materially capable of supporting a great deal of display-oriented activity and that is nevertheless unable to correct itself.

The film *Minority Report*, loosely based on a 1956 short story by Philip K. Dick, offers a suggestive image of this requirement [8, 33]. As the premise is usually summarized, specially gifted persons, the precogs, produce visions of future crimes on which an institution acts, and because there are several of them their visions can differ. The record of a dissenting vision is the minority report of the title. In the film, such differing visions are purged from the records, which preserves an appearance of infallibility [33]. An institution that acts on such outputs has reason to present them as unanimous, since its authority rests on appearing reliable.

The image can be read as an analogy for any arrangement in which a small class of specialists produces predictions or estimates on which an organization acts. Quantitative analysts who model risk, and programmers who build the systems that classify and score, may be thought of as occupying such positions in relation to the businesses and institutions that employ them. The analogy captures a position and not a person. Unlike the precogs, analysts are paid, can leave, and can dissent in public, and the comparison should not be taken to suggest otherwise. What it indicates is a recurring structure. The organization consumes a signal from a few producers, the credibility of the signal rests on its appearing singular and confident, and the alternative models, the stated uncertainties and the dissenting estimates, which are the minority reports of the arrangement, are the first things to be discarded when they complicate a decision. Curtis argues in the interview discussed above that the concept of risk came to organize managerial and financial life from the 1990s onward, and that this orientation tends toward the preservation of stability [4]. If that interpretation is accepted, the pressure to present risk estimates as settled is a particular case of the pressure to reproduce a stable form described in the preceding section.

The premise also illustrates the difference between redundancy and independence. Several sources provide protection only if their failure mechanisms differ. Sources that share a failure mechanism fail together, and the appearance of several concordant judgments can mislead. Nor does the preservation of a dissent establish that the dissent is correct. The defensible principle is that potentially relevant disagreement must remain available for evaluation, together with procedures for assessing its significance. A dissenting record that exists but is

not permitted to inform decisions, or that is treated automatically as malfunction, fails the conditions of informed assessment and route to intervention stated in this essay. A dissenting record that is accepted automatically fails differently, since it substitutes one unexamined authority for another.

The second requirement of viability concerns the diversity of modes of work. Societies depend heavily on roles that formalize, model and optimize, such as programming, statistics and economic analysis, and they have good reason to value those roles and to extend authority to those who perform them. It is plausible that part of the deference such roles attract derives from this dependence, since the society knows that its systems of provision, finance and communication cannot run without them. Deference of this kind is not unreasonable, but it carries two risks. The first is that formalization handles only what can be made legible, in the sense developed in Scott's account [29]. The care, tacit skill and local repair that formal systems presuppose are performed by others and go unrecorded, so that a society composed predominantly of formalizing roles would lack the people who supply what the formal layer omits. The second is that a population engaged in similar modes of work shares error modes. The blind spots of the formal layer would then be common to nearly everyone placed to notice them, which weakens the capacity for correction. A diversity of working methods can reduce such shared blind spots, although it does not establish a maximum permissible share of formalizing roles. Deference also bears on the minority report, since the standing granted to a dependable source of estimates makes a dissenting estimate harder to hear.

An earlier essay by the present author argues in a parallel way about exploratory cognition. A society in which every citizen behaved as a researcher would deplete its reserves of order, so such cognition must remain selective, a minority trait supported by institutional structure and by the maintenance and care of others, within a spectrum of cognitive roles whose diversity contributes to resilience [11]. The same structure applies here, with an important qualification. The requirement concerns roles and modes of work and not kinds of people. The same person may formalize in one setting and care or repair in another, and nothing in the argument implies that those who formalize are lesser or unusual. The requirement yields no proportion. It says that a society's allocation across modes of work is itself a variable to be kept within a viable range, and that high regard for the most formalizable roles can obscure how much depends on the others.

## **10. Pleasure, Leisure, and the Objection from Paternalism**

The strongest objection to the argument is that it licenses paternalism. Who decides which satisfactions are legitimate? People may reasonably prefer entertainment, companionship or

immersion to activities that an outside observer considers more substantial, and they are entitled to do so. An argument against wireheading that became an argument that everyone must remain productive would be both unattractive and self-undermining, since it would reproduce at the level of a moral theory the very instrumentalism it set out to criticize. Leisure is not the opposite of a good life, and much that is most worth having consists of activities pursued for their own sake.

The response is to locate the concern elsewhere than in the content of anyone's preferences. The relevant considerations are agency, dependency and displacement. Can people reconsider their participation in a given arrangement, and do they have access to the information and alternatives that reconsideration requires? Do they retain meaningful alternatives, or have the alternatives been eroded by the growth of the arrangement itself? Are the conditions of future choice being preserved? And is the arrangement supported by costs that participants cannot see, or that others must bear? These questions permit criticism without treating pleasure as suspect. They take a form familiar from capability approaches in political philosophy, which assess arrangements by the real freedoms they leave people to choose and to revise their choices, and not by the particular choices people make [30].

Nozick's experience machine is instructive in this respect. The standard reading holds that people decline to plug in because they value contact with reality. A more modest reading, adequate for the present purposes, is that people value the retention of the capacity to leave and to be corrected by what lies outside the experience. An arrangement that provides pleasant experience while preserving that capacity differs importantly from one that forecloses it. On this view, a person who spends an evening absorbed in a film is not wireheaded, and neither is a community that enjoys shared entertainment. A system becomes a candidate for the diagnosis when its design makes reconsideration difficult, when it depends on concealed costs, or when it erodes the material and social conditions under which alternatives could be pursued.

## 11. A Universalizability Test for Attention Strategies

The argument of the preceding sections can be given a deontological formulation. The question is whether the principles organizing an activity could be endorsed as general terms of cooperation. Kant's formula of universal law asks whether the maxim of an action could be willed as a universal law without contradiction or without frustrating the agent's own ends [18], and Rawlsian contractualism asks whether principles could be justified to all those affected under conditions of fair reasoning [23]. Applied to attention strategies, the test discriminates between two maxims. "Communicate what you can contribute" can be adopted by everyone without defeating itself, because its success for one person does not depend

on others failing to follow it. “Continually capture more attention than your competitors” is different in kind. Everyone can follow it, and some will continue to win, so it is not self-defeating in the way that a contradiction would be. What cannot be universalized is everyone’s achieving superior relative standing, since not everyone can rank above others. The stronger criticism concerns two further features: the collective costs of escalation, which rise as each participant matches the efforts of the others, and the presentation of positional success as a general solution, which misleads those who cannot all achieve it. These considerations support the ethical argument without making competitive conduct as such a contradiction under universalization.

The universalization argument must be qualified with care. A society requires differentiated roles, and an occupation need not be suitable for everyone in order to be legitimate. Not everyone can be a surgeon, a farmer or a conductor, and the fact that a role cannot be universalized is not an objection to it. The distinction that matters is between a role whose existence is compatible with a fair general arrangement, and a position whose attractiveness depends on concealing its reliance on others. The relevant question is whether the governing arrangement can sustain the different roles fairly, with the less visible work recognized and supported, rather than presenting a position dependent on others’ hidden labour as everyone’s route to freedom.

Put differently, the test applies to the *form of the advice* and not to the occupation. An adviser who says that some people can make a living through public communication, and describes honestly what that depends upon, offers something that can be universalized as information. An adviser who says that any individual can try, and who treats this as showing that everyone can succeed simultaneously, offers something that works only so long as most listeners do not act on it fully. The latter is not thereby shown to be insincere. It is shown to be a prescription that cannot serve as a general term of cooperation.

## 12. Preserving the Possibility of Correction

The positive proposal that emerges from the analysis is that rewards should be kept answerable to their purposes. This is a modest-sounding aim with demanding implications. Institutions need ways to discover when their preferred measurements cease to track useful outcomes, and people need room to pursue work whose value appears slowly or resists continuous publicity. The aim is not to find a better metric that will finally measure everything correctly. The analysis does not show that every targeted metric must eventually drift, only that optimization can undermine a proxy under identifiable conditions, so that no metric should be assumed immune. It is to preserve the capacity of a system to detect and respond to such drift.

Several design questions follow, and each involves trade-offs that should be stated frankly. The first concerns protected time for learning and maintenance. Arrangements that insulate a portion of effort from continuous measurement allow work to proceed whose value is not yet legible, but they also create room for idleness and for the persistence of unproductive projects, and distinguishing the two is itself a judgment that metrics cannot make. The second concerns support for long research horizons. Funding structures with long commitments allow investigators to pursue questions that do not yield a steady stream of outputs, at the cost of reduced short-term accountability and the risk of entrenchment. The third concerns opportunities to leave engagement systems. Exit, in Hirschman’s sense, supplies a signal about quality that the system cannot easily manufacture [17], but making exit easy may reduce the revenue of the systems that offer it, and the incentive to resist is therefore structural. The fourth concerns forms of recognition that do not require permanent self-advertisement. Mechanisms of peer recognition, apprenticeship lineages, and records of maintained work can supply acknowledgement of contributions without demanding a continuous public presence, but they depend on communities with enough stability and enough shared standards to administer them.

An additional resource lies in the practice of making the supporting work visible in a form that is not performance. The didactic design discussed earlier is one instance: a system that displays its own preparation and repair makes available to observers some of what it takes to sustain the performance and where it fails. Documentation of ordinary maintenance, public repair records, and open accounts of failed attempts serve a similar function in institutions. They are valuable not because they attract attention but because they keep the difference between the signal and the activity open to inspection. Russell’s suggestion that the stories of maintainers be told points to a related use of publicity, which can help people recognize neglected contributions and secure support for them [28]. Visibility becomes a problem only when the demands of presentation displace the work, or when recognition never translates into better conditions for those who do it.

Observation alone does not amount to correction. Visibility must support informed assessment, and informed assessment must have an effective route to intervention. Three conditions can accordingly be distinguished. Failure must be *observable*, so that the discrepancy can be seen. Observers must be able to make an *informed assessment*, which requires that they can understand what they see and diagnose its cause. And informed assessment must have an *effective route to intervention*, so that contrary evidence can alter what the system does. The third condition is the most demanding, since it requires that people can formulate alternatives, coordinate their efforts, and alter the arrangements they recognize as defective. Widespread acknowledgement of a defect without any of these capacities leaves the defect in place. A robot can display its repair process without observers

being able to diagnose or alter it, and an institution can publish its failures while ignoring them. Each measure discussed above can be assessed against all three conditions, and a measure that secures only the first may serve chiefly as a further favourable indicator. These conditions enable correction. They do not guarantee that every proposed correction will be sound, since a route to intervention can carry mistaken proposals as readily as sound ones, which is a further reason why the assessment condition matters.

A fifth question follows from the discussion of hypernormalization, and concerns the conversion of private knowledge into public information. Where participants understand that a measure has drifted but face a cost for saying so, the relevant design problem is to lower that cost or to route the information around it. Anonymous reporting, protected dissent in review processes, independent audit with real standing, and the routine publication of negative results are examples of arrangements with this aim, and each has familiar weaknesses: anonymity can shelter bad faith, protection can be hollow in practice, and audit can itself become a performed ritual of the kind it was meant to detect. The test of such an arrangement is whether it changes what participants would be willing to say in public, and not whether it produces a further favourable indicator.

None of these measures is offered as a solution. Each can itself be gamed, and each introduces costs. The claim is the more limited one that a system which has no such measures is more exposed to the failure described, since nothing within it can register that a failure has occurred.

### **13. A World Worth Attending To**

The argument may be summarized by returning to the difference between securing attention and sustaining its objects. A functioning society needs experiences that people enjoy, and it needs institutions that recognize achievement and communicate it. It also needs the material, ecological, intellectual and interpersonal conditions under which enjoyment and recognition remain possible. These conditions are produced largely by activities that attract little notice while they succeed. A culture that rewards the capture of attention more reliably than it rewards the maintenance of what attention is directed toward is likely to underinvest in precisely those conditions, and its members may be the last to notice, because the reinforcement they receive will remain strong.

The deepest danger of wireheading, understood in this extended sense, is not that people take pleasure in what they do. It is the progressive loss of the capacity to notice and repair the difference between feeling rewarded and living well. When that capacity is intact, a society can enjoy its entertainments, pursue its ambitions and communicate freely, and correct its course when its satisfactions begin to diverge from its needs. When it is lost, the

divergence widens without producing any internal alarm. A common world must therefore remain available to correct our satisfactions, even when that correction is inconvenient, unprofitable or initially unrewarding. The condition for its remaining available is that some of what people attend to, and some of what they do, continues to be answerable to something other than the signal it produces. The principle can be stated in three parts: failure must be observable, observable failure must be open to informed assessment, and informed assessment must have an effective route to intervention, which presupposes that alternatives can be formulated and collective action coordinated. A common world remains corrective only when encountering contrary evidence can actually alter what the system does. These conditions enable correction and do not guarantee that each proposed correction is sound.

## References

- [1] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.
- [2] Argall, B. D., Chernova, S., Veloso, M., and Browning, B. (2009). A survey of robot learning from demonstration. *Robotics and Autonomous Systems*, 57(5), 469–483.
- [3] Berridge, K. C., and Robinson, T. E. (1998). What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? *Brain Research Reviews*, 28(3), 309–369.
- [4] Brand, R. (host), and Curtis, A. (guest). (2017). Do we really want change? *Under the Skin with Russell Brand*, episode 3 [podcast; video posted 22 July 2017 on the host’s YouTube channel]. Cited from an auto-generated transcript.
- [5] Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning*, 2(1), 67–90.
- [6] Curtis, A. (director). (2016). *HyperNormalisation*. BBC documentary film.
- [7] Debord, G. (1967). *La Société du spectacle*. Paris: Buchet-Chastel.
- [8] Dick, P. K. (1956). The minority report. *Fantastic Universe*, 1956.
- [9] Everitt, T., and Hutter, M. (2016). Avoiding wireheading with value reinforcement learning. In *Artificial General Intelligence (AGI 2016)*, Lecture Notes in Computer Science 9782, 12–22. Springer.
- [10] Felton, S., Tolley, M., Demaine, E., Rus, D., and Wood, R. (2014). A method for building self-folding machines. *Science*, 345(6197), 644–646.

- [11] Flyxion. (2025). The ecology of neoteny: institutions, entropy, and the steady-state civilisation. Essay, November 2025. <https://standardgalactic.github.io/alphabet/The%20Ecology%20of%20Neoteny.pdf>
- [12] Frank, R. H., and Cook, P. J. (1995). *The Winner-Take-All Society*. New York: Free Press.
- [13] Goodhart, C. A. E. (1975). Problems of monetary management: the U.K. experience. In *Papers in Monetary Economics*, vol. 1. Sydney: Reserve Bank of Australia.
- [14] Granovetter, M. (1978). Threshold models of collective behavior. *American Journal of Sociology*, 83(6), 1420–1443.
- [15] Graeber, D. (2018). *Bullshit Jobs: A Theory*. New York: Simon and Schuster.
- [16] Hirsch, F. (1976). *Social Limits to Growth*. Cambridge, MA: Harvard University Press.
- [17] Hirschman, A. O. (1970). *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States*. Cambridge, MA: Harvard University Press.
- [18] Kant, I. (1785). *Grundlegung zur Metaphysik der Sitten*. Riga: Hartknoch.
- [19] Kuran, T. (1995). *Private Truths, Public Lies: The Social Consequences of Preference Falsification*. Cambridge, MA: Harvard University Press.
- [20] Muller, J. Z. (2018). *The Tyranny of Metrics*. Princeton, NJ: Princeton University Press.
- [21] Nozick, R. (1974). *Anarchy, State, and Utopia*. New York: Basic Books.
- [22] Olds, J., and Milner, P. (1954). Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. *Journal of Comparative and Physiological Psychology*, 47(6), 419–427.
- [23] Rawls, J. (1971). *A Theory of Justice*. Cambridge, MA: Harvard University Press.
- [24] Ring, M., and Orseau, L. (2011). Delusion, survival, and intelligent agents. In *Artificial General Intelligence (AGI 2011)*, Lecture Notes in Computer Science 6830, 11–20. Springer.
- [25] Rosen, S. (1981). The economics of superstars. *American Economic Review*, 71(5), 845–858.
- [26] Rus, D., and Tolley, M. T. (2015). Design, fabrication and control of soft robots. *Nature*, 521, 467–475.

- [27] Russell, A., and Vinsel, L. (2016). Hail the maintainers. *Aeon*, 6 April 2016.
- [28] Russell, A. (interviewee), and Schlemmer, R. (interviewer). (n.d.). Is innovation the enemy of maintenance? *The Agility Effect* (VINCI Energies). <https://www.theagilityeffect.com/en/about/innovation-enemy-maintenance/>. The page shows no publication date.
- [29] Scott, J. C. (1998). *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. New Haven, CT: Yale University Press.
- [30] Sen, A. (1999). *Development as Freedom*. New York: Knopf.
- [31] Simon, H. A. (1971). Designing organizations for an information-rich world. In M. Greenberger (ed.), *Computers, Communications, and the Public Interest*, 37–72. Baltimore: Johns Hopkins Press.
- [32] Skalse, J., Howe, N., Krasheninnikov, D., and Krueger, D. (2022). Defining and characterizing reward gaming. In *Advances in Neural Information Processing Systems 35*. (Circulated on arXiv as “Defining and characterizing reward hacking,” arXiv:2209.13085.)
- [33] Spielberg, S. (director). (2002). *Minority Report*. DreamWorks Pictures and Twentieth Century Fox.
- [34] Star, S. L. (1999). The ethnography of infrastructure. *American Behavioral Scientist*, 43(3), 377–391.
- [35] Strathern, M. (1997). “Improving ratings”: audit in the British university system. *European Review*, 5(3), 305–321.
- [36] Ukeles, M. L. (1969). *Manifesto for Maintenance Art 1969!* Proposal for an exhibition, “Care.” (Excerpted in *Artforum*, 1971.)
- [37] Wu, T. (2016). *The Attention Merchants: The Epic Scramble to Get Inside Our Heads*. New York: Knopf.
- [38] Yurchak, A. (2006). *Everything Was Forever, Until It Was No More: The Last Soviet Generation*. Princeton, NJ: Princeton University Press.