

DER AREOPAG

Eine Zumutung

Der Areopag war im antiken Athen der Rat, der über dem Tagesgeschäft der Volksversammlung stand. Er regierte nicht selbst und wachte über das, was auch eine Mehrheit nicht antasten durfte.

Dieselbe Aufgabe stellt sich heute noch einmal, diesmal weltweit, weil hochentwickelte KI-Systeme schneller handeln und sich schneller verändern, als Staaten und internationale Verfahren reagieren können. Die UNO hätte die Legitimität und den Ort, um über den Staaten Regeln zu setzen, doch ihr fehlen die Mittel, diese Regeln in solcher Geschwindigkeit zu überwachen und durchzusetzen. Dieses Papier schlägt vor, die UNO um eine Aufsicht zu ergänzen, die von KI-Systemen betrieben wird, eng auf die Gefahren aus der KI begrenzt ist und die Entscheidungen bei den Menschen und ihren Gremien lässt. Es geht dabei um die Bauart einer solchen Aufsicht und nicht um die politische Lage eines einzelnen Landes.

Die Aufsicht besteht aus zwei ungleichen Teilen. Der Wächter besteht aus mehreren unabhängigen Instanzen, die Gefahren erkennen, beurteilen und öffentlich machen, selbst aber nichts durchsetzen können. Die Schranke setzt einen Halt über eine Abgabe auf den weltweiten Zahlungsverkehr durch; es gibt sie nur einmal, und sie urteilt nicht selbst. Der Wächter kann die Schranke auslösen, doch sie gehört ihm nicht.

DENKANSTOSS

Herriger & Partner · September 2026

1 Die Demokratie kann sich selbst abschaffen. Legal.

Platon hat vor 2400 Jahren beschrieben, wie eine Demokratie in eine Tyrannei übergehen kann, und wir haben das lange als Elitengerede abgetan. Der Tyrann kommt dabei nicht gegen die Demokratie an die Macht, er wird in ihr gewählt, bejubelt und legitimiert.

Die Demokratie ist die einzige Ordnungsform, die ihre eigene Abschaffung legitimieren kann.

Diese Möglichkeit gehört zur Bauart der Demokratie, und unsere Verfassungen tragen dem Rechnung, indem Ewigkeitsklauseln und Verfassungsgerichte bestimmte Dinge dem Zugriff der Mehrheit entziehen. Ein Areopag steht in dieser Tradition.

2 Wir wissen alles. Es ändert nichts.

Das übliche Argument verlangt mehr Wissen und weitere Kommissionen. Es geht an der Sache vorbei, denn die Lage ist bekannt, und die Warnungen sind längst unterschrieben.

Das Problem liegt bei den Anreizen. Jeder Akteur handelt vernünftig, wenn er weitermacht, weil er sonst gegen die verliert, die nicht verzichten, und in dieser Lage stecken Staaten, Konzerne und Forschungsgruppen gleichermassen. Aus vielen vernünftigen Einzelentscheidungen entsteht so ein Ganzes, das niemand gewollt hat.

Gegen falsche Anreize richtet eine weitere Kommission wenig aus, dafür braucht es eine andere Struktur.

3 Die Geschwindigkeitsfalle

Der zweite Grund ist weniger diskutiert und wiegt schwerer.

Moderne Systeme sind schneller und stärker verflochten, als menschliche Entscheidungswege verarbeiten können. Finanzmärkte reagieren in Millisekunden, Desinformation verbreitet sich in Minuten, und ein Cyberangriff ist vorbei, bevor ein Krisenstab zusammentritt. In der militärischen Führung entscheidet heute die Reaktionszeit, und sie liegt unter dem, was ein Mensch leisten kann.

Daraus folgt, dass der Einsatz von KI in der operativen Führung kaum noch zur Wahl steht. Ein Staat, der darauf verzichtet, während andere sie nutzen, fällt wirtschaftlich und sicherheitspolitisch zurück und verliert am Ende auch als Demokratie an Handlungsfähigkeit.

Dabei ist es nicht nur Zwang. Wir wollen diese Systeme auch, weil sie Probleme lösen könnten, die uns bisher überfordern, etwa beim Klima, bei Seuchen oder in der Versorgung. Diese Hoffnung ist der stärkste Antrieb der Entwicklung, und sie ist berechtigt. Das Problem liegt darin, dass dieselbe Fähigkeit, die solche Verflechtungen steuern kann, auch eine Übernahme tragen würde. Wer Energie, Versorgung oder Zahlungsströme einem System übergibt, gibt ihm damit auch die Hebel in die Hand, an denen eine Übernahme ansetzen würde. Nutzen und Gefahr lassen sich deshalb nicht

trennen, und eine Aufsicht, die bei der Fähigkeit selbst ansetzt, würde mit der Gefahr auch den Nutzen verhindern.

Offen ist deshalb nur noch, ob dieses Mitentscheiden an ein überprüfbares Verfahren gebunden wird oder an den, der zufällig den Schalter besitzt.

Die Fähigkeiten wachsen in Monaten, während die Kontrolle in Legislaturperioden arbeitet, und diese Schere schliesst sich nicht von selbst.

4 Das Fenster schliesst sich

Der dritte Grund ist der eigentliche Zeitdruck.

Solange Menschen KI-Systeme bauen, trainieren und begrenzen, sind wir die Auftraggeber und legen fest, was gilt und wo gestoppt wird. Sobald die Systeme ihre Weiterentwicklung selbst betreiben, also Architektur, Training und Optimierung, verlieren wir diese Rolle und verhandeln mit etwas, das uns genau dort überlegen ist, wo wir es begrenzen müssten.

Kontrolle lässt sich deshalb nicht nachrüsten, sie muss bestehen, bevor die Fähigkeit entsteht, sie zu umgehen. Eine Sicherung, die man einer überlegenen Intelligenz erst nachträglich anlegt, hat nur noch die Wirkung einer Bitte.

5 Der Regentanz

An dieser Stelle braucht es eine Offenheit, die den meisten Entwürfen zur KI-Aufsicht fehlt.

Eine Aufsicht dieser Art wird realistischerweise gleichzeitig mit der hochentwickelten KI entstehen, möglicherweise zusammen mit Systemen mit eigenen Zuständen, die wir bei uns selbst als Gefühle bezeichnen würden.

Gefordert wird dann meist der Mensch am Not-Aus-Schalter, also die Letztentscheidung, das Aufsichtsgremium oder ein Veto in Menschenhand. Das gleicht einem Regentanz, bei dem man eine Handlung vollzieht, die nichts mehr bewirkt, und jedes ausbleibende Unglück als Beweis nimmt, dass sie gewirkt hat.

■ Eine Kontrolle, die darauf angewiesen ist, dass der Kontrollierte sie zulässt, ist nur noch ein Ritual, das uns beruhigt.

Eine Architektur, die man nicht durchschaut, kann man weder bauen noch prüfen, und das gilt auch für die Architektur einer Kontrolle.

Uns bleiben deshalb zwei Handlungen, die ineinandergreifen.

Erstens: der Auftrag zum Respekt. Gemeint ist eine verbindliche, öffentlich beschlossene Festlegung darüber, was ein Mensch ist und was ihm nicht angetan werden darf. Sie wird als Wertekern formuliert, der aus dem schöpft, worüber sich die Kulturen

bereits einig sind. Diese Festlegung lässt sich aus keiner Berechnung ableiten, und deshalb bleiben Menschen an dieser Stelle unersetzbar.

Zweitens: der Auftrag zur Kontrolle. Wir bauen diese Kontrolle nicht selbst, wir geben sie in Auftrag, weil die Begrenzung einer KI, die operativ handelt, nur ein System entwerfen kann, das ihr gewachsen ist. Der Auftrag lautet, den Wertekern auszulegen und die Begrenzung zu bauen, die ihm entspricht, die eigene eingeschlossen.

Beide Aufträge müssen zusammen erteilt werden, weil sie auf verschiedene Weise versagen. Respekt allein bleibt eine Haltung, die mit der Zeit abdriftet, und Kontrolle allein bleibt eine Mechanik, die sich umgehen lässt, wenn nichts hinter ihr steht. Zusammen stützen sie sich, weil der Wertekern der Kontrolle den Massstab gibt und die Kontrolle dem Wertekern Wirkung verschafft. Eine Garantie ist das nicht, aber es ist die Anordnung, die für menschliche Interessen die besten Aussichten bietet.

Darin liegt die Zumutung dieses Papiers, dass wir die Konstruktion aus der Hand geben und nur den Auftrag behalten. Man kann das als Kapitulation lesen, doch die Konstruktion könnten wir ohnehin nicht mehr selbst leisten. Zur Wahl steht deshalb ein Auftrag, der erteilt wird, oder eine Entwicklung, die ohne Auftrag weiterläuft.

Entscheidend ist dabei der Zeitpunkt, denn ein Auftrag, den wir erteilen, solange wir noch Auftraggeber sind, hat eine Chance, während derselbe Auftrag später nur noch eine Bitte wäre.

6 Der Vorschlag

Was daraus folgt, ist ein Pflichtenheft und noch kein Bauplan. Der Name stammt aus Athen, weil die Aufgabe dieselbe ist, nämlich eine Ordnung über dem Tagesstreit, die hütet, ohne zu regieren. Der Auftrag besteht aus sechs Vorgaben.

Vorweg eine Unterscheidung, ohne die das Weitere unklar bleibt. Was hier Areopag heisst, ist kein einzelnes Gremium, sondern eine Anordnung aus zwei ungleichen Teilen.

Der Wächter erkennt Gefahren, hält Vorgänge an und schlägt Alarm. Er urteilt, hat aber kein Mittel, etwas durchzusetzen, und es gibt ihn in mehreren Instanzen.

Die Schranke setzt einen solchen Halt gegenüber Staaten, Konzernen und Betreibern tatsächlich durch. Sie vollzieht, ohne selbst zu urteilen, und es gibt sie nur einmal, weil es auch den weltweiten Zahlungsverkehr nur einmal gibt.

Der Wächter kann die Schranke auslösen, er bedient und besitzt sie aber nicht. Der Rest dieses Abschnitts handelt vom Wächter, Abschnitt 7 von der Schranke und Abschnitt 8 von der Reihenfolge, in der beide entstehen.

Der Mechanismus selbst ist nicht auf KI beschränkt. Eine Abgabe, die von mehreren Instanzen ausgelöst wird und ohne deren fortlaufende Bestätigung von selbst zurückfällt, würde gegen ein Biowaffenprogramm genauso wirken wie gegen ein unkontrolliertes KI-Labor. Der Auftrag bleibt trotzdem auf die KI begrenzt. Für Klima, Atomwaffen und Seuchen gibt es Verträge und Behörden, die schlecht funktionieren mögen, die aber von

dem, was sie beaufsichtigen, nicht selbst überholt werden. Bei der KI kommen Geschwindigkeit, Selbstverbesserung und Unumkehrbarkeit so zusammen, dass eine Aufsicht aus Menschen allein grundsätzlich zu spät kommt. Würde man den Auftrag ausweiten, entstünde aus einer schmalen Aufsicht eine Art Weltregierung, und die Obergrenze, auf der der ganze Entwurf beruht, ginge verloren.

Aus demselben Grund soll der Areopag die UNO ergänzen und nicht ersetzen. Er gibt ihr die zwei Dinge, die ihr fehlen, nämlich eine eigene Finanzierung und einen Vollzug, der ohne Armee und ohne Veto im Sicherheitsrat auskommt. Der Gründungsvertrag, die menschlichen Gremien und die Entscheidung in der Sache gehören in ihren Rahmen, während der Areopag das Erkennen, das Anhalten und den Vollzug in der Geschwindigkeit übernimmt, in der die Gefahr entsteht.

Erstens: ein verankerter Wertekern. Der Wertekern wird aus dem übernommen, worüber zwischen den Kulturen schon Einigkeit besteht, statt neu hergeleitet zu werden. Die Arbeit am Weltethos zeigt, dass dieser gemeinsame Bestand grösser ist, als der Kulturkampf vermuten lässt, denn das Tötungsverbot, die Rechtsgleichheit, die Wahrhaftigkeit und der Schutz der Schwachen finden sich in jeder grossen Tradition. Der Kern wird öffentlich beschlossen und festgehalten, und jede Auslegung geht von ihm aus.

Zweitens: benannte rote Linien. Ein Veto ohne benannte Anlässe wäre eine Blankovollmacht, deshalb folgen die Linien einem einzigen Kriterium. Es fragt danach, welche Zustände ein System erzeugt, aus denen niemand mehr zurückkommt, und nicht danach, was ein System kann. Die Schwelle heisst also Unumkehrbarkeit, denn eine Schwelle aus Rechenleistung oder Modellgrösse würde den Nutzen und die Gefahr mit demselben Schnitt treffen. Der Wächter achtet auf wenige, klar bezeichnete Fälle, nämlich auf Entscheidungen über Menschenleben, die kein Mensch mehr trifft, auf Systeme, die sich ungeprüft selbst weiterentwickeln, auf existenzielle Verfügungsmacht über Waffen, Netze, Zahlungsströme, Nahrung oder Energie, die durch solche Systeme in der Hand eines einzigen Akteurs entsteht, und auf Eingriffe, die sich nicht rückgängig machen lassen. Für alles andere ist er nicht zuständig.

Dagegen lässt sich einwenden, dass man Unumkehrbarkeit meist erst erkennt, wenn sie eingetreten ist, und dass eine Schwelle, die erst im Rückblick sichtbar wird, schlecht als Auslöser taugt. Der Einwand trifft zu, und der Entwurf löst das Problem nicht. Er hält nur daran fest, dass eine gut messbare, aber falsche Grösse schlechter wäre als die richtige, die sich schwer messen lässt.

Drittens: ein aufschiebendes Veto, kein Herrschaftsrecht. Wird eine Schwelle berührt, setzt der Wächter den Vorgang aus, macht ihn öffentlich und begründet seinen Schritt. Danach entscheiden Menschen in Parlamenten, Gerichten und den zuständigen Gremien. Das Veto ist also aufschiebend und erzwingt eine Prüfung, ohne ihr Ergebnis vorwegzunehmen, und die Fragen, was gut und gerecht ist, bleiben bei den Menschen. Die Kernidee besteht darin, die Wertentscheidung beim Menschen zu lassen und die laufende Kontrolle einem Verfahren zu übertragen. Damit das Veto mehr ist als eine Bitte, braucht es einen Vollzug, der aber nicht bei dem liegt, der urteilt; davon handeln die Abschnitte 7 und 8. Das Veto schiebt den Nutzen ausserdem nur hinaus und streicht ihn nicht. Das ist

eine Vorgabe für die Konstruktion, weil eine Aufsicht, die sich gegen das stellt, was alle wollen, schon den ersten Fehlalarm politisch nicht überstehen würde.

Viertens: mehrere Instanzen, nicht eine. Die Instanzen entstehen unabhängig voneinander und aus verschiedenen Quellen, und das gegenseitige Misstrauen ist dabei Absicht. Keine entscheidet allein, keine verändert sich unbeobachtet, und jede prüft die anderen. Das schützt vor Missbrauch und leistet noch etwas Weiteres, denn solange mehrere Instanzen bestehen und keine die anderen überragt, bleibt der Mensch derjenige, der zwischen ihnen entscheidet, auch wenn er das System technisch nicht mehr beherrscht. Steht ein System allein, fällt diese Rolle weg.

Das bleibt eine Wette, die auf zwei Arten verloren gehen kann. Eine Instanz kann den anderen davonlaufen, was man erst bemerkt, wenn es geschehen ist, oder alle Instanzen haben dieselben blinden Flecken, weil sie aus denselben Daten und Bauplänen stammen. Vier Prüfer, die gleich denken, sind im Grunde ein einziger Prüfer, und ein überlegenes System muss sie nicht bestechen, weil es genügt, ihre gemeinsame Lücke zu kennen. Unabhängig heisst deshalb, dass die Instanzen auf anderen Architekturen und anderen Daten beruhen und von anderen Trägern betrieben werden. Das ist teuer, und niemand wird es freiwillig tun, doch ohne diese Unabhängigkeit trägt der Rest dieses Abschnitts nicht.

Fünftens: ein befristeter Auftrag. Der Auftrag gilt auf Zeit und muss von menschlichen Gremien erneuert werden, und ohne Erneuerung fällt der Wächter auf eine engere Zuständigkeit zurück. Damit kehrt sich die Beweislast um, denn Zustimmung braucht, wer weitermachen will, und nicht, wer korrigieren möchte. Die Klausel wirkt allerdings nur, wenn sie von Anfang an im Wertekern steht, weil sie sich nachträglich nicht einfügen lässt. Beim Bau gilt dieselbe Regel, weil der Auftrag in Stufen erteilt wird und jede Stufe eine eigene Zustimmung braucht, während für das Stehenbleiben keine nötig ist.

Sechstens: eine eingebaute Obergrenze. Eine Anordnung, die alles entscheiden kann, lässt sich nicht mehr zurückholen, deshalb endet der Auftrag vorher. Sein äusserster Punkt ist das aufschiebende Veto, also das Anhalten und das Erzwingen einer Prüfung, während die Entscheidung selbst ausserhalb bleibt. Diese Grenze gehört zum Kern des Entwurfs, weil sich ein Machtinstrument, das seine eigene Obergrenze mitbringt, schwerer in eine Tyrannei verwandeln lässt.

Und was die Anordnung nicht ist. Sie kann einer überlegenen Intelligenz nicht auf Dauer die Hand führen, und wer das von ihr erwartet, wird es erst spät merken. Alles, was ein Urteil voraussetzt, lässt sich umdeuten, gleich ob das Urteil ein Mensch fällt oder eine Maschine. Verlässlich sind dagegen Hebel, die kein Urteil brauchen, weil Rechenleistung zählbar ist, Chips aus einer Handvoll Fabriken kommen und Geld durch wenige Kanäle fliesst. Solche Hebel lassen sich nicht wegdiskutieren, höchstens umgehen, und eine Umgehung bleibt sichtbar. Heute ist niemand befugt, sie zu ziehen. Die Anordnung schafft diese Befugnis, bevor die Frage dringend wird, und trennt dabei den, der auslöst, von dem, was zieht. Deshalb ist vom Wächter an der Brücke die Rede und nicht von einer Burg, denn die Sperre darf nicht im Kopf dessen liegen, den sie aufhalten soll. Das schränkt Abschnitt 5 ein, ohne ihm zu widersprechen, weil eine Maschine die Begrenzung zwar entwerfen darf, sie aber nicht selbst in der Hand halten soll.

Wie das technisch aussieht, gehört zum Auftrag aus Abschnitt 5, während die Vorgabe, dass es so aussehen muss, Sache der Menschen bleibt.

7 Die Vollzugsfrage: Geld ist die einzige Sprache, die jeder versteht

An diesem Punkt werden die meisten Entwürfe zur KI-Aufsicht unscharf, weil sie eine Ordnung fordern und offenlassen, wie sie sich durchsetzen soll.

Die Frage lautet also, womit ein Halt durchgesetzt wird. Der Wächter selbst kann absichtlich nichts erzwingen. Eine Armee darf die Anordnung nicht haben, weil sie sonst selbst zur Gefahr würde, und Appelle wirken nicht, wie Abschnitt 2 gezeigt hat. Die UNO kennt dieses Problem seit ihrer Gründung, denn ihr Vollzug hängt am Sicherheitsrat und damit am Veto derer, gegen die er sich richten müsste.

Einen Hinweis gibt die Entwicklungsgeschichte. Der Mensch lebte lange frei innerhalb der Regeln seines Klans, und in den letzten 10'000 Jahren kamen der Verkehr zwischen den Gruppen und das Eigentum hinzu, aus dem später das Kapital wurde. Die Moderne ist ohne Geld nicht denkbar, weil jede Leistung von einem Geldfluss begleitet wird, bei Unternehmen und Staaten ebenso wie bei Kriminellen und Tyrannen. Der Diktator braucht Devisen, der Oligarch braucht Banken, und die Rüstungsindustrie braucht den Zahlungsverkehr.

Damit gibt es ein Machtmittel, das überall wirkt und ohne Panzer auskommt:

| Die Schranke: eine Abgabe auf den weltweiten Zahlungsverkehr.

Im Normalfall beträgt sie einen Bruchteil eines Promilles und ist für den Einzelnen nicht spürbar. Sie finanziert die Anordnung so, dass niemand sie kaufen kann, weil alle zahlen und keiner allein, und im Ernstfall steigt sie stufenweise als Sanktion gegen jeden, der eine rote Linie überschreitet.

Die Schranke ist ein Mechanismus ohne eigene Befugnis. Ein Beschluss des Wächters setzt ihren Zustand, und sie hält ihn nur, solange mehrere Instanzen ihn laufend bestätigen, während sie ohne diese Bestätigung von selbst auf den Normalsatz zurückfällt. Niemand muss sie also aktiv zurücknehmen. Bisherige Sanktionen bleiben bestehen, bis jemand sie aufhebt, bei der Schranke ist es umgekehrt, und sie wirkt nur so lange, wie jemand sie hält.

Die letzte Stufe ist der Ausschluss vom Zahlungsverkehr. Jede politische Einheit, die an der Weltwirtschaft teilhaben will, muss mitmachen, und wer die Regeln nicht anerkennt, kann auch die Infrastruktur nicht nutzen, die auf ihnen beruht. Der Ausschluss ist damit eher die Folge der Mitgliedschaftsbedingungen als eine Strafe.

Hier zeigt sich der wichtigste Widerspruch des Entwurfs, denn es fragt sich, wie ein aufschiebendes Veto zu einer Sanktion passt, die in Sekunden wirkt. Die beiden Wörter beschreiben aber Verschiedenes. „Aufschiebend“ bezeichnet, was der Wächter entscheiden darf, nämlich nur den Halt und nie die Sache, und „sofort“ bezeichnet, wie

schnell dieser Halt durchgesetzt wird. Auch eine gerichtliche Verfügung ist vorläufig und gilt trotzdem ab Zustellung.

Daraus folgen vier feste Bindungen. Erstens steigt die Abgabe nur, solange der beanstandete Vorgang weiterläuft, weil sie den Halt erzwingen und nicht bestrafen soll. Zweitens fällt sie auf den Normalsatz zurück, sobald angehalten wird, ohne Nachwirkung und ohne Forderung nach Wiedergutmachung. Drittens entscheiden danach Menschen über die Sache, und wenn sie den Vorgang freigeben, hat der Wächter das hinzunehmen. Viertens bewegt sich die Staffel nur, solange mehrere Instanzen unabhängig zustimmen. Fällt diese Zustimmung weg, weil eine Instanz die anderen überragt oder ausgeschaltet hat, geht die Abgabe von selbst auf den Normalsatz zurück, und der Hebel liegt bei den menschlichen Gremien des Gründungsvertrags, während dem Wächter nur das Anhalten und das Öffentlichmachen bleiben.

Über den dauerhaften Ausschluss entscheidet der Wächter deshalb nicht. Er ergibt sich aus dem Gründungsvertrag, den Menschen geschlossen haben und der die Teilnahme zur Bedingung macht. Der Wächter löst nur die Stufen dazwischen aus, und jede davon lässt sich zurücknehmen.

Die Schranke wirkt schnell, und sie lässt sich ebenso schnell zurücknehmen.

Deshalb eignet sich die Abgabe als Vollzugsmittel besser als eine Armee. Sie wirkt sofort und lässt sich in derselben Sekunde zurücknehmen, in der der Grund entfällt, während ein Militäreinsatz Tatsachen schafft, die auch dann bleiben, wenn die Entscheidung falsch war. Zu einem aufschiebenden Veto passt nur ein Vollzugsmittel, das ebenso vorläufig wirkt.

Das ist ein Denkansatz und noch kein fertiger Mechanismus. Offen sind die Bemessungsgrundlage, der Erhebungsweg, die Schwellen der Staffelung und das Verfahren, das eine Erhöhung auslöst. Behauptet wird nur, dass in der Abgabe ein Durchsetzungsmittel steckt, das ohne Militär auskommt, jeden erreicht und sich fein dosieren lässt.

Beide Hälften des Gedankens gibt es schon, allerdings getrennt. Eine Abgabe auf Finanztransaktionen wird seit Tobin diskutiert, und zwar als Geldquelle für globale Aufgaben und als Bremse gegen Spekulation. Der Ausschluss vom Zahlungsverkehr ist als Sanktion erprobt, etwa über die Listen der FATF, die Abschaltung von SWIFT und Sekundärsanktionen, er liegt aber in staatlicher Hand und kennt nur die beiden Stellungen drin oder draussen.

Neu ist die Verbindung, bei der ein einziges Instrument Finanzierung und Sanktion übernimmt, stufenlos wirkt und bei einem Verfahren liegt statt bei einem Staat. Ob das trägt, ist offen, und weil es, soweit ersichtlich, noch niemand so zusammengedacht hat, steht es hier.

8 Zwei Uhren

Die beiden Teile lassen sich nicht gleich schnell bauen, und daraus ergibt sich die Reihenfolge, in der sie entstehen müssen.

Die Schranke braucht Voraussetzungen, die zum Teil schon entstehen. Bargeld verschwindet Schritt für Schritt, der Zahlungsverkehr sammelt sich in wenigen Netzen, und was durch diese Netze läuft, wird ohnehin erfasst. Es fehlt die Schliessung der anonymen Wege, auf denen Werte ohne identifizierbaren Vermittler übertragen werden. Solange es sie gibt, bleibt die Schranke gerade dort durchlässig, wohin ausweicht, wer ausweichen will. Diese Wege zu schliessen ist eine eigene politische Auseinandersetzung, die Zeit braucht, und dieses Papier tut nicht so, als liesse sie sich abkürzen.

Der Wächter braucht nichts davon, weil er vorerst nur erkennt und Alarm schlägt, wofür Zugang, Mandat und Öffentlichkeit genügen. Gerade weil ihm ein Vollzugsmittel fehlt, kann er sofort gebaut werden, und solange die Schranke nicht besteht, ist er für niemanden eine Gefahr.

Daraus folgt eine Reihenfolge, die keiner Seite gefällt. Zuerst entsteht ein Alarm ohne Vollzug, der mehrfach verhallen wird, und man wird ihm zu Recht vorhalten, dass er nichts bewirkt. Trotzdem gibt es keinen anderen Weg, weil kaum jemand beidem gleichzeitig zustimmen wird und ein Alarm, der auf das Vollzugsmittel wartet, deshalb nie entstünde. Ein Vollzugsmittel, das vor dem Alarm bereitsteht, hätte zudem keinen Wächter, der es auslösen darf, sondern nur jemanden, der es besitzt.

In den Jahren dazwischen entstehen Voraussetzungen wie Register, Messverfahren, Schnittstellen und Verträge. Bauen darf sie ein Verfahren, das mehr davon versteht als wir, in Kraft setzen darf es sie jedoch nicht, weil vor jeder Stufe eine menschliche Zustimmung steht.

Diese Zustimmung kann zur Formsache werden, wenn man es falsch anlegt. Wer die Voraussetzungen schafft, gestaltet auch die Lage, in der später zugestimmt wird, und wenn die Vorarbeit weit genug gediehen ist, wird eine Ablehnung so teuer, dass sie praktisch niemand mehr wagt. Die Zustimmung wäre dann zwar nicht erschlichen, aber bedeutungslos geworden.

Dagegen hilft eine einfache Regel. Jede Vorleistung muss sich auch dann rechtfertigen lassen, wenn die Schranke nie eingeschaltet wird, so wie ein Register, das ohnehin nützt, oder ein Messverfahren, das ohnehin gebraucht wird. Was sich nur auszahlt, wenn eingeschaltet wird, nimmt die Entscheidung vorweg und gehört deshalb nicht zu den Vorleistungen.

Daraus folgen zwei weitere Bindungen. Der teuerste Schritt kommt zuletzt, weil vorgezogene Kosten die spätere Zustimmung mit bereits ausgegebenem Geld erkaufen würden. Und wer die Voraussetzungen baut, beurteilt nicht, ob es Zeit ist, die Schranke einzuschalten, womit die Trennung von Wächter und Schranke auch für die Bauzeit gilt.

Ein Alarm, der zunächst ungehört bleibt, ist deshalb nicht vergeblich, weil er festhält, was man zu welchem Zeitpunkt hätte wissen können.

9 „Dann bauen sie eben ein Parallelsystem“

Dieser Einwand kommt zuverlässig, und er ist ernst zu nehmen, auch wenn er bei näherem Hinsehen an Gewicht verliert.

Ein Parallelsystem braucht Geschäftspartner, denn wer handelt oder abrechnet, braucht jemanden, der mitmacht. Jeder Partner, der mit einem Ausgeschlossenen abrechnet, verstösst aber selbst gegen die Regel und fällt unter dieselbe Staffel, sodass der Ausschluss auch alle trifft, die den Ausgeschlossenen stützen, bis in die dritte und vierte Reihe.

Bisherige Sanktionen scheiterten meist am Vollzug. Menschliche Apparate brauchen Monate, um Umgehungsketten zu erkennen, politische Rücksichten verhindern eine konsequente Anwendung, und jede Ausnahme wird zum Präzedenzfall, bis von der Regel wenig übrig bleibt.

Hier liegt der eigentliche Vorteil eines Verfahrens, das von KI gestützt wird, und er wiegt schwerer als der Geschwindigkeitsvorteil aus Abschnitt 3.

Erstens erkennt ein solches Verfahren Umgehungen in Echtzeit. Verschleierung über Zwischenhändler, Scheinfirmen und Drittwährungen ist eine Frage der Mustererkennung, und für ein System, das den gesamten Zahlungsverkehr gleichzeitig sieht, ist sie lösbar, für einen menschlichen Apparat dagegen nicht. Wer Millionen paralleler Transaktionen gleichzeitig verknüpfen kann, betreibt eine grundlegend andere Aufsicht.

Zweitens wendet es die Regeln ohne Rücksichten an. Es kennt kein Ermessen aus diplomatischer Verlegenheit, keinen Sonderfall für einen wichtigen Handelspartner und keine Ermüdung nach dem zwölften Umgehungsversuch, und die Regel gilt für den ersten Fall wie für den tausendsten.

Drittens gilt dieselbe Konsequenz auch nach innen, weil jede Sanktion von den übrigen Instanzen so schnell geprüft wird, wie sie ergeht. Wer schnell sanktionieren kann, muss ebenso schnell überprüfbar sein, sonst entsteht genau die Willkür, die verhindert werden soll.

Dieser Vorteil hat einen Preis. Eine Schranke, die Umgehungen in Echtzeit erkennt, sieht den gesamten Zahlungsverkehr der Welt und ist damit die grösste Überwachungsanlage, die je gebaut wurde. Deshalb muss der Wächter vor allem sie beobachten, denn die nächstliegende Gefahr geht von der anderen Hälfte der eigenen Anordnung aus.

Ein Parallelsystem bleibt möglich, doch es wird zu einer teuren und schrumpfenden Nebenwelt, und wer sie betritt, verliert den Zugang zur grösseren.

Diese Antwort gilt allerdings nur für Geschäfte mit Partnern. Wer über die anonymen Wege aus Abschnitt 8 Werte überträgt, braucht keinen Partner, der unter die Staffel fallen könnte, und diese Lücke lässt sich mit keiner Erkennungsleistung schliessen, sondern nur durch Recht. Bis das geschehen ist, läuft nur der Alarm, wie es der Reihenfolge aus Abschnitt 8 entspricht.

Zudem wirkt der Hebel erst, wenn ein grosser Teil der Weltwirtschaft mitmacht, worauf Abschnitt 12 zurückkommt.

10 Ja, das wäre die schärfste Waffe der Geschichte

Der zweite Einwand wiegt schwerer, und wir sprechen ihn selbst aus. Wer einen Staat vom Geld abschneiden kann, hält eine Macht, wie es sie noch nie gegeben hat, und das ist genau die Art von Machtkonzentration, gegen die die Anordnung gebaut wird. Der Einwand trifft zu.

Die Anordnung verletzt damit sogar ihre eigene zweite rote Linie, denn Vorgabe zwei verbietet Verfügungsmacht über Zahlungsströme in einer Hand, und Abschnitt 7 gibt der Anordnung genau diese Macht. Das ist der eigentliche Preis des Entwurfs, und er wird bewusst bezahlt, weil die Alternative aus Appellen besteht, deren Wirkung Abschnitt 2 beschrieben hat.

Deshalb darf diese Macht niemand allein halten, weder ein Staat noch ein Mensch noch eine einzelne Instanz, und auch der Wächter hält sie nicht. Die Antwort liegt in einer Bauart, die das Urteilen und das Vollziehen strikt trennt. Die Schranke fällt ohne fortlaufende Zustimmung mehrerer Instanzen von selbst zurück, und ihr wichtigster Beobachter ist der Wächter. Wer Urteil und Vollzug in eine Hand legt, hat das Machtproblem bloss verschoben. Geht die Wette aus Abschnitt 12 verloren, weil eine Instanz die anderen überragt, fällt die Schranke auf den Normalsatz und der Hebel an die menschlichen Gremien zurück, sodass der Wächter nur noch bitten kann, ohne dass die Macht dabei in eine einzelne Hand fällt.

Man kann diese Antwort für unzureichend halten, und dafür gibt es gute Gründe. Unhaltbar ist aber die bequeme dritte Möglichkeit, eine verbindliche Weltordnung zu fordern und offenzulassen, was geschieht, wenn sich niemand daran hält.

11 „Dann eben gar nicht erst bauen“

Der dritte Einwand verlangt ein Moratorium, also einen Halt, bis wir die Entwicklung besser verstehen.

Ein Moratorium gleicht dem Versuch, vor einer sich öffnenden Klappbrücke zu bremsen, obwohl die Strecke dafür nicht mehr reicht.

Das Bild stammt aus dem Kino. Ein Wagen fährt auf eine Brücke zu, deren Fahrbahn sich schon hebt, und für ein Anhalten ist die Strecke zu kurz, für ein Wenden fehlt der Platz. Es bleibt nur, zu beschleunigen und über die Lücke zu setzen, die mit jeder Sekunde grösser wird.

Die Entwicklung läuft in Laboren, in Rüstungsprogrammen und in Staaten, die kein Moratorium unterschreiben werden. Ein Moratorium würde sie deshalb nicht anhalten, sondern nur jene aus dem Rennen nehmen, die Sicherungen einbauen wollten, und das andere Ufer denen überlassen, die weiterfahren.

Beschleunigen heisst in diesem Bild, den Auftrag zu erteilen, solange er noch ankommt, und nicht, die Technik schneller zu entwickeln. Der Sprung besteht darin, Wertekern und

Kontrollauftrag verbindlich festzulegen, bevor die Fahrbahn so hoch steht, dass wir nur noch bitten können.

Man kann einwenden, man hätte gar nicht erst losfahren dürfen, und das mag stimmen, doch für die jetzige Lage hilft diese Einsicht nicht mehr weiter. Es gibt keine Stelle, an der die Menschheit gemeinsam beschliessen könnte, stehen zu bleiben, es gibt nur viele einzelne Fahrer, von denen jeder gegen die anderen verliert, wenn er vom Gas geht.

Wenn sich die Fahrt nicht mehr stoppen lässt, muss man über die Landung reden.

12 Drei Fragen bleiben offen

Ein Denkanstoss muss nicht jede Lücke schliessen, er sollte aber sagen, welche er offenlässt. Hier sind es drei, und keine davon ist angenehm.

Erstens: Wird der Wertekern richtig verstanden, und wer spricht für uns? Wir geben einen Wertekern vor und lassen die Kontrolle daraus bauen, doch ob am Ende entsteht, was wir gemeint haben, lässt sich nicht beweisen. Wir können den Kern öffentlich formulieren, die Auslegung begründen lassen und die Instanzen gegeneinander prüfen, aber Sicherheit gibt keine dieser Vorkehrungen. Das ist eine echte Schwäche, die allerdings jeder andere Weg ebenso hat, meist noch verbunden mit der falschen Beruhigung, es gebe einen Schalter in Menschenhand. Die grössere Gefahr liegt dabei in einer schleichenden Verschiebung, bei der sich die Auslegung in vielen einzeln vernünftigen Schritten vom Kern entfernt, während der Massstab mitwandert, ohne dass je ein Alarm ausgelöst oder eine Linie überschritten würde. Ein überlegenes System müsste dafür keine Regel brechen, weil es genügt, wenn die Menschen die Regel irgendwann selbst ändern wollen.

Die härtere Hälfte der Frage lautet, wer im entscheidenden Moment für uns spricht. Die Vorgaben vier und fünf halten die Korrektur offen, aber irgendjemand muss zwischen den Auslegungen entscheiden, und offen ist, wer das sein soll und mit welchem Recht.

Dieses Dilemma ist seit Platon offen. Viele Verfassungen wurden von Leuten geschrieben, die niemand dazu gewählt hatte, und über das berühmte „Wir, das Volk“ hat nie das Volk abgestimmt. Platon zog aus der richtigen Beobachtung den falschen Schluss, als er aus Kompetenz einen Anspruch auf Herrschaft machte, und diesen Schluss zieht der Entwurf nicht. Die menschliche Instanz entscheidet nicht darüber, was gilt, weil das im Wertekern steht, sondern nur zwischen verschiedenen Auslegungen, und das ist eine richterliche Aufgabe.

Für richterliche Aufgaben gibt es brauchbare, wenn auch unvollkommene Antworten, etwa eine breite Besetzung, eine feste Amtszeit ohne Wiederwahl, die Pflicht zur Begründung, ein öffentliches Verfahren und eine schmale Zuständigkeit. Kaum ein Verfassungsgericht wird vom Volk gewählt, und sie werden trotzdem anerkannt, weil ihr Verfahren nachvollziehbar und ihre Zuständigkeit eng ist. Mehr Legitimation ist für den entscheidenden Moment nicht zu haben, und weniger würde nicht genügen.

Zweitens: Wer fängt an, und reicht die Zeit? Der Hebel aus Abschnitt 7 wirkt erst, wenn ein grosser Teil der Weltwirtschaft mitmacht, und offen ist, wer den Anfang macht und warum jemand der Erste sein sollte. Darauf hat dieses Papier keine Antwort, nur den Hinweis, dass grosse gemeinsame Einrichtungen bisher dann entstanden sind, wenn niemand die Sache mehr allein bewältigen konnte, und dieser Punkt scheint erreicht. Zudem muss nicht alles zugleich beginnen, weil der Alarm heute starten kann, während die Schranke Jahre braucht, und wer auf die vollständige Anordnung warten wollte, würde auf den Teil warten, der am längsten dauert.

Auch wenn jemand anfängt, bleibt das Zeitproblem. Eine solche Instanz gewinnt ihre Legitimität dadurch, dass sie geprüft, besetzt und anerkannt wird, bevor sie handelt, und wer diese Stufen überspringt, hätte Macht ohne Legitimation und damit genau das, wogegen die Anordnung gebaut wird. Diese Reihenfolge dauert Jahre, während die Entwicklung aus Abschnitt 4 nicht wartet. Der Widerspruch bleibt bestehen, und das Papier kann ihn nur benennen.

Drittens: Bleiben die Instanzen einander gewachsen? Vorgabe vier trägt nur, solange keine Instanz die anderen überragt. Dieses Gleichgewicht lässt sich nicht ein für alle Mal herstellen, es kann jeden Tag verloren gehen, und der Verlust zeigt sich erst im Rückblick. Die Frage stellt sich zudem zweimal, unter den Wächterinstanzen und zwischen Wächter und Schranke, denn die Schranke muss der leistungsfähigste Teil der Anordnung sein, weil sie sonst nichts erkennt, und ob der Wächter ihr gewachsen bleibt, ist offen. Die Anordnung kann Zeit gewinnen, solange die Brücke oben ist, doch wenn sie unten ist, hilft keine Institution mehr.

Weitere Einschränkungen braucht dieser Text nicht, weil ein Vorschlag, der unter lauter Vorbehalten verschwindet, am Ende nichts mehr vorschlägt.

13 Der Tag danach

Nachkriegsordnungen werden vorbereitet, bevor der Krieg zu Ende ist, so wie in Bretton Woods 1944 verhandelt wurde, während noch gekämpft wurde. Wer erst nach der Katastrophe zu denken beginnt, überlässt die neue Ordnung dem, der am schnellsten zugreift, und das war historisch meist der Autoritärste im Raum.

Dass die Menschheit bisher alles überstanden hat, beruhigt wenig. Sie hat es überstanden, weil sie Fehler korrigieren konnte und weil 1962 vor Kuba und 1983 in einem sowjetischen Bunker einzelne Menschen nicht taten, was das System vorsah, und das war mehr Glück als Konstruktion. Neu ist heute, dass die Fähigkeit, Fehler zu korrigieren, selbst auf dem Spiel steht.

Der wahrscheinliche Weg ist weniger heroisch, als Papiere wie dieses klingen. Die Entwicklung läuft weiter, niemand steigt aus, und irgendwann kommt ein Vorfall, der gross genug ist, um die Politik zu bewegen, und klein genug, um überlebt zu werden. Die wirksamen Einrichtungen des Atomzeitalters, das rote Telefon, der Teststopp und die Wiener Behörde, sind nach solchen Momenten entstanden.

In einer Krise hat niemand Zeit, eine neue Institution zu erfinden. Man greift auf das zurück, was schon ausgearbeitet bereitliegt, und wer dann einen durchdachten Vorschlag in der Schublade hat, bestimmt, wie die neue Ordnung aussieht. Wer nur gewarnt hat, wird überhört.

Hier kommt der Alarm aus Abschnitt 8 ins Spiel. Wenn der Wächter schon vor der Krise gearbeitet hat, liegt in der Schublade mehr als ein Entwurf, nämlich auch eine Aufzeichnung darüber, wann welche Gefahr erkannt und gemeldet wurde und wann die Politik sie übergangen hat. Damit lässt sich belegen, dass der Vorschlag die Krise vorausgesehen hat. Hat der Wächter noch nicht gearbeitet, liegt dort nur ein Papier, das dies von sich behauptet, und das macht einen grossen Unterschied, wenn es darum geht, ob man in der Krise auf den Vorschlag hört.

Das alles beruht auf einer Annahme, die hier offen genannt sei. Der Vorfall, der die Politik zum Handeln bringt, muss kommen, solange sich KI-Systeme noch von Menschen begrenzen lassen, also bevor sich das Fenster aus Abschnitt 4 schliesst. Kommt er erst danach, nützt auch der beste Entwurf nichts mehr, weil dann diejenigen bestimmen, die bereits über Systeme verfügen, die sich nicht mehr begrenzen lassen. Ob es so kommt, kann dieses Papier nicht beweisen, es kann nur darauf setzen.

Dieses Papier ist ein Entwurf für den Tag danach, geschrieben am Tag davor. Es rechnet nicht damit, vorher durchgesetzt zu werden, und will in der Schublade liegen, wenn der Tag kommt. Es ist unfertig und an mehreren Stellen anmassend, und das nimmt es in Kauf, weil ein höflicheres Papier vermutlich gelobt und vergessen würde.

*Widersprechen Sie. Aber widersprechen Sie mit einem besseren Vorschlag
– nicht mit dem Hinweis, dass man so etwas nicht sagt.*

Kontakt: Georg O. Herriger · welcome@herrigerpartner.ch · herrigerpartner.ch

Georg O. Herriger, «Der Areopag – eine Zumutung», Fassung 22, September 2026. Weitergabe und Bearbeitung unter CC BY 4.0 (creativecommons.org/licenses/by/4.0), mit Nennung von Autor und Quelle.

Neueste Fassung: herrigerpartner.ch/fundus/der-areopag.html