On ne gouverne que ce dont on peut encore modifier la trajectoire

La gouvernabilité ne s'inventorie pas, elle s'éprouve sous perturbation et elle est distincte de la résilience.

1. Ce qui décide n'est pas la présence du contrôle, mais sa performance sous perturbation

Quand l'Arizona impose, à compter du 1er juillet 2026, qu'un directeur médical signe personnellement tout refus de prise en charge produit par un algorithme (HB 2175), et que le Colorado fait de même au 30 juin (SB 24-205), le législateur obtient exactement ce qu'il vise : ***de l'imputabilité***.

Une signature désigne un responsable, ouvre un recours, fonde une contestation. Ce sont des objectifs légitimes, et la signature les sert. Ce qu'elle ne sert pas, c'est la gouvernabilité du dispositif. Elle atteste qu'un humain était présent ; elle n'atteste rien sur la capacité du système à rattraper la décision si elle se révèle fausse trois semaines plus tard, quand le patient est déjà sorti du circuit. La signature est nécessaire à l'imputabilité. Elle est insuffisante à la gouvernabilité. Les deux ne se confondent pas.

C'est le point aveugle du discours sur l'humain dans la boucle. On évalue un dispositif de supervision à sa capacité de corriger les erreurs qu'on attendait. On ne l'évalue presque jamais à ce qui décide réellement du sort des systèmes : leur comportement face à ce qu'ils n'avaient pas prévu. Les erreurs connues occupent le quotidien. Les perturbations non prévues redessinent les systèmes.

La thèse de cette note tient en une phrase, et elle est plus générale que toute mécanique particulière : ***la gouvernabilité d'un dispositif ne se démontre pas par la présence de mécanismes de sûreté, mais par leur performance lorsqu'on les éprouve***. Un coupe-circuit qui existe sans avoir jamais été actionné n'est pas une garantie, c'est une décoration. De cette thèse découle une condition concrète, la capacité de récupération, (dont nous verrons qu'elle prend deux formes) et qu'elle ne se confond ni avec la gouvernabilité qu'elle conditionne, ni avec la résilience qu'elle subsume.

Domaine de validité, et un avertissement de niveau. La démonstration porte sur les systèmes automatisés dotés d'un appareil de supervision identifiable. Elle ne prétend pas couvrir les ordres distribués sans superviseur central (marchés, communs numériques, écosystèmes ouverts…), qui relèvent d'une autre analyse. Surtout, elle

emprunte une partie de ses contre-exemples à l'aviation, aux réseaux électriques, à la conduite de crise. Ce déplacement, de la supervision algorithmique vers les systèmes sociotechniques critiques, n'est pas un glissement subi : il est revendiqué. Ces systèmes sont les instances mûres de la même relation de supervision sous perturbation, et c'est à ce titre, et à ce titre seulement, qu'ils servent de banc d'essai. La note parle donc des deux à la fois , supervision IA et systèmes complexes, mais d'un seul objet : ***la relation entre un dispositif automatisé et ce qui le gouverne***. Là où l'argument vaut pour le second sans valoir pour le premier, ce sera dit.

2. Trois mots qu'on confond : résilience, récupération, gouvernabilité

Le mot qui porte la démonstration doit être défini avant d'être employé, sous peine de devenir un attracteur sémantique qui absorbe tous les problèmes. Mais une définition isolée ne suffit pas : le risque, ici, n'est pas le flou d'un terme, c'est la confusion de trois. Posons-les comme une chaîne, du système vers son superviseur.

La *résilience* est une propriété du système : sa capacité à absorber une perturbation sans intervention extérieure (redondance, confinement, dégradation progressive, marges…). ***Un système résilient encaisse seul***.

La *récupération* est une propriété opératoire, et c'est elle qui doit recevoir un critère observable sous peine de devenir le nouveau mot-valise. Je l'appelle la capacité à restaurer un espace acceptable d'états après le franchissement d'une perturbation non représentée. Trois objets y sont nommés et donc mesurables :

1. Un espace d'états jugé acceptable,
2. Un événement qui en sort,
3. Un retour dans le domaine acceptable.

La récupération se décline en deux formes, on y vient ; mais elle n'est pas « tout ce qui aide un système à survivre ». Elle est ce mouvement précis : restaurer, après sortie, un domaine défini. Par exemple, l’activation d’un Plan de Reprise d’Activité (PRA) après sinistre (évènement qui sort de l’espace des états jugés acceptables).

La *gouvernabilité*, enfin, n'est pas une propriété du système. C'est une propriété de la *relation* entre le système et son superviseur : la capacité de ce dernier à maintenir le système dans un espace acceptable, ou à l'y ramener, attestée à l'épreuve de la perturbation et à travers les régimes. Le poids est sur « à l'épreuve ». La gouvernabilité se constate sous charge, elle ne se déduit pas d'un inventaire de mécanismes.

Cette chaîne tranche d'un coup deux objections qui, sinon, ruineraient la note.

1. La première est l'objection de tautologie : si la récupération englobe rollback, résilience, confinement, redondance, dégradation et réparation, alors « gouvernable » finit par vouloir dire « capable de continuer à être gouverné », ce qui est vrai et circulaire. La parade n'est pas de rétrécir la récupération mais de la *situer* : la récupération est ce que le système rend disponible ; la gouvernabilité est ce que le superviseur parvient à en mobiliser, et qu'on a éprouvé. Un système peut être hautement récupérable et mal gouvernable, parce que le superviseur n'a pas la main sur ses leviers de retour, ne les voit pas, ou n'ose pas les actionner. La récupération est nécessaire à la gouvernabilité ; elle n'en est pas la définition.
2. La seconde objection est plus sérieuse. Un lecteur familier de Perrow, de Hollnagel ou de Woods dira que tout ceci reformule la résilience des systèmes complexes sous un vocabulaire neuf. L'objection serait décisive si la gouvernabilité était une propriété du système car elle ne serait alors qu'un synonyme tardif de résilience. Mais elle est une propriété de la relation système/superviseur, et c'est exactement ce que la littérature de la résilience laisse dans l'ombre : un système peut être intrinsèquement résilient et pourtant échapper à son superviseur, et un système peu résilient rester gouvernable parce qu'un superviseur informé peut le rattraper à temps.
    a. La résilience décrit ce que le système fait seul.
    b. La gouvernabilité décrit ce qu'un tiers peut encore en faire.

    La résilience est une vertu du système ; la gouvernabilité est une vertu du couple. C'est cette différence d'objet, et non un mot nouveau, qui fait la contribution.

Disposer de ressources de gouvernance n'est d'ailleurs pas la même chose qu'être gouvernable. L'armée française de 1940 disposait de l'observation, de l'autorité hiérarchique, des moyens d'intervention. Elle s'est effondrée parce que ces ressources n'étaient pas articulées au rythme de l'événement. À l'inverse, des opérateurs de réseaux ont tenu des systèmes critiques avec des moyens rudimentaires, parce que leur articulation était juste. La causalité « moyens, donc gouvernabilité » est fausse. Ce qui se mesure, c'est la tenue des moyens sous contrainte, pas leur nombre.

Trois distinctions, enfin, évitent de confondre la gouvernabilité avec ses voisines immédiates.

1. Elle n'est pas l'*alignement* : un dispositif parfaitement gouvernable peut servir une fin destructrice, il est neutre sur l'objet.
2. Elle n'est pas la *pilotabilité instantanée* : un pilote désengage le pilote automatique, il ne réécrit pas les lois de commande en vol ; agir *dans* le système et agir *sur* le système sont deux pouvoirs distincts.
3. Elle n'est pas la *contrôlabilité locale* : un système incontrôlable point par point peut rester gouvernable dans son ensemble. Mais la séparation d'avec

l'alignement, utile, n'est pas étanche : on ne contrôle que relativement à une finalité, et des objectifs contradictoires rendent un dispositif ingouvernable même richement doté, parce que l'autorité ne sait plus au nom de quoi intervenir.

3. La récupération, et ses deux formes

Si la gouvernabilité est une performance sous perturbation, la propriété du système qui la rend possible est la capacité de récupération. La précision du §2 est ici essentielle, car la récupération n'a de sens que rapportée à un seuil et à des variables. Un refus de remboursement peut être juridiquement réversible quand le dommage clinique qu'il a causé ne l'est plus biologiquement. Ce qui compte n'est pas la réversibilité en général, c'est la récupération des variables qui décident de la survie, et avant le point de non-retour.

Cette capacité prend deux formes, qu'il faut séparer parce qu'on les confond.

1. La première est la *réversibilité* : revenir en arrière, défaire. C'est la forme la plus puissante quand elle est disponible, et elle se décline elle-même en plusieurs registres qu'on ne doit pas tenir pour équivalents : réversibilité technique (rollback d'un système), décisionnelle (annulation d'une décision), juridique (voie de recours), physique ou clinique (réparation d'un dommage réel). Un dispositif peut être réversible dans un registre et pas dans un autre, et c'est précisément là que se logent les illusions de sûreté : on certifie un rollback technique en laissant intact un dommage physique déjà advenu.
2. La seconde forme est la *résilience* : survivre sans retour arrière. Beaucoup de systèmes critiques fonctionnent sur des décisions largement irréversibles (chirurgie, aviation, gestion de crise nucléaire, conduite de guerre…). Ils restent pourtant gouvernables, non parce qu'ils peuvent revenir en arrière, mais parce qu'ils absorbent le dommage. La formulation juste n'est donc pas que la réversibilité est le socle de tout, mais que la réversibilité est la forme la plus puissante de la récupération, et que là où elle est impossible, la résilience en tient lieu. La capacité de récupération est nécessaire ; la réversibilité en est le cas favorable, pas l'unique modalité.

On notera le rapport exact avec le §2 : la résilience apparaît ici comme l'une des deux formes que peut prendre la récupération, côté système, et non comme un concept rival. Ce que la gouvernabilité ajoute, c'est la question que ni la réversibilité ni la résilience ne posent d'elles-mêmes : ce retour ou cette absorption sont-ils à la portée d'un superviseur, au bon moment, et l'a-t-on vérifié ?

4. Pourquoi certains systèmes absorbent les surprises et d'autres s'effondrent

Une démonstration qui ne s'appuie que sur des catastrophes prouve peu. Three Mile Island, Challenger, la faillite de Long-Term Capital Management, le Flash Crash de 2010, Fukushima : la liste est saisissante, mais on n'y trouve que des systèmes où la perturbation a gagné. On peut dresser la liste symétrique, celle des systèmes qui absorbent l'imprévu sans s'effondrer : le contrôle aérien encaisse chaque jour des situations hors procédure, les réseaux électriques gèrent en continu des défaillances locales, l'aviation civile moderne a fait de l'incident non prévu un cas traité plutôt qu'un destin. Une théorie sérieuse doit expliquer les deux listes, pas seulement la première.

Ce qui sépare les deux listes n'est pas la chance, c'est la capacité de récupération mobilisable. Les systèmes qui tiennent ont construit, en amont, de quoi absorber ou défaire : redondance du contrôle aérien, confinement et délestage des réseaux, marges et procédures de récupération de l'aviation. Les systèmes qui s'effondrent avaient, au moment décisif, franchi un seuil irréversible sans capacité de retour ni d'absorption. Les catastrophes ne réfutent donc pas la thèse, elles l'instancient : ce sont les cas où la récupération manquait.

Encore faut-il stabiliser le mot autour duquel tout tourne. « Surprise » est trop vague : il désigne tantôt l'événement improbable, tantôt l'inconnu, tantôt la dérive lente. La catégorie pertinente est plus précise : une ***perturbation non représentée dans les hypothèses de conception***. Ce qui compte n'est pas qu'un événement soit rare ou spectaculaire, mais qu'il tombe hors de ce que le dispositif a prévu de traiter. Une dérive lente de modèle et un choc brutal appartiennent à la même catégorie dès lors qu'aucun des deux ne figurait dans les hypothèses, et la première est souvent plus dangereuse que le second, parce qu'elle ne déclenche aucune alarme.

Mais cette définition par la non-représentation se paie d'une difficulté qu'il faut affronter, non contourner. Une perturbation non représentée ne s'identifie comme telle qu'*après coup* : ***tant qu'elle n'a pas eu lieu, elle est, par construction, hors du champ de ce qu'on sait nommer***. Comment, dès lors, démontrer *ex ante* qu'un dispositif saura traiter ce qui, par définition, n'est pas encore connu ? La réponse honnête est qu'on ne le peut pas, et qu'il faut renoncer à le prétendre. On ne démontre jamais la capacité à gérer une perturbation spécifique inconnue. On démontre seulement la présence, et la performance, sous épreuve, de *capacités génériques de récupération* : largeur de l'espace d'états restaurable, rapidité du retour, indépendance des chemins de rattrapage, marge avant l'irréversible. Ces capacités s'exercent sur des perturbations représentées, qui tiennent lieu de procuration pour celles qui ne le sont pas. La procuration est imparfaite, et c'est une limite irréductible, pas un détail : éprouver une capacité de récupération sur le connu reste la seule donnée disponible sur son comportement face à

l'inconnu. La gouvernabilité ne supprime pas l'incertitude sur l'inédit ; elle déplace le pari, de « avons-nous prévu cet événement ? » vers « avons-nous une machine de rattrapage assez générique, et l'avons-nous fait tourner ? ». C'est un meilleur pari. Ce n'est pas une preuve d'invulnérabilité.

5. Les ressources sont un graphe, pas une pyramide

Il serait commode de ranger les conditions de la récupération en une hiérarchie nette. Ce serait trop élégant. La capacité de récupération s'appuie sur quatre ressources qui interagissent plus qu'elles ne s'empilent :

1. l'*observabilité* (voir ce que fait le système),
2. l'*intelligibilité* (comprendre pourquoi),
3. l'*autorité* (pouvoir agir sans sanction asymétrique),
4. la *capacité d'intervention* (disposer des moyens d'agir).

Leur dépendance n'est pas linéaire mais croisée.

Sans observabilité suffisante, la récupération devient inutilisable : on ne défait pas ce qu'on ne voit pas, et un coupe-circuit qu'on ne sait pas quand actionner ne protège de rien. Mais une forte observabilité peut compenser une intelligibilité faible : on pilote la vapeur de Watt, les antibiotiques, le deep learning bien au-delà de ce qu'on en explique, parce qu'on en observe les effets et qu'on garde la main. Et sans autorité, l'intervention est fictive : un superviseur qu'on sanctionne pour avoir bloqué un flux apprend à ne plus bloquer.

Le système ressemble à un graphe de dépendances, pas à une pyramide, et il faut résister à la tentation de transformer cette heuristique en ontologie. On remarquera que ces quatre ressources sont précisément les variables de la *relation* définie au §2 : elles ne décrivent pas le système seul, mais ce qu'un superviseur peut en voir, en comprendre, en décider et y faire. C'est pourquoi elles relèvent de la gouvernabilité et non de la seule résilience. L'intérêt n'est pas de les classer, il est de poser à un système concret quatre questions opposables, plus une cinquième qui les commande toutes : que peut-il récupérer, et l'a-t-on éprouvé ?

6. Pourquoi les dispositifs réels perdent leur capacité de récupération

Si la récupération est la condition, comment expliquer qu'autant de dispositifs s'en privent sans que personne ne l'ait décidé ? C'est ici, et non dans la taxonomie des ressources, que se loge la contribution la plus féconde : passer de « ce qui est nécessaire » à « pourquoi cela disparaît ». Le §5 dresse l'inventaire statique de ce qu'il faudrait ; le

présent paragraphe en donne la dynamique d'érosion. Quatre mécanismes font, ensemble, du décor un équilibre rationnel, c'est-à-dire un état vers lequel des acteurs raisonnables convergent sans que nul ne l'ait voulu.

1. Le premier est *économique* : l'automatisation effondre le coût marginal de la décision, l'examen sérieux garde le sien. La revue approfondie, qui coûtait peu rapportée à un faible volume, cesse d'être rentable rapportée à un volume devenu massif. Ce n'est pas la rigueur qui devient impossible, c'est son ratio coût/bénéfice qui s'inverse.
2. Le deuxième est *organisationnel*, et c'est le plus tenace : la responsabilité est asymétrique. On sanctionne le superviseur qui a bloqué un flux légitime. L'erreur est visible, datée, imputable ; on sanctionne rarement celui qui a laissé passer une erreur de la machine car la faute se dilue dans le système. Le superviseur apprend, sans qu'on le lui dise, que valider est sans risque et bloquer en comporte un. Il valide. C'est un comportement optimal pour lui, et destructeur pour la récupération.
3. Le troisième est *cognitif* : à mesure que le modèle ne se trompe pas, la vigilance s'érode par habitude. La validation, d'abord acte réfléchi, devient réflexe : on entérine les choix/décisions du système parce qu'il n'a jamais commis d'erreurs grossières jusqu'à présent. Le superviseur reste présent, la signature du §1 est bien là, mais sa présence a cessé de produire de l'examen. C'est la complaisance d'automatisation, et elle est d'autant plus forte que le modèle est *bon* : un système médiocre maintiendrait la garde par ses ratés mêmes.
4. Le quatrième est *propre à l'agentique*, et c'est le plus récent : quand un système orchestre lui-même planification, délégation et exécution, l'enchaînement des décisions cesse d'être lisible. Plus personne ne sait pourquoi telle action a eu lieu, et le pouvoir de défaire porte alors sur un objet devenu opaque. On peut conserver l'autorité et les moyens, et perdre malgré tout la récupération, faute d'intelligibilité du chemin à rebrousser.

Ces quatre mécanismes ne s'additionnent pas, ils se renforcent : l'économie supprime le temps de l'examen, l'organisation en supprime l'incitation, la cognition en supprime l'habitude, l'agentique en supprime l'objet. La récupération ne disparaît pas par décision, elle s'érode par dilution. Aucun de ces mécanismes ne suppose une intention. Il serait donc faux d'écrire que l'IA transforme le contrôle en théâtre, comme si quelqu'un en avait monté la scène. L'IA déplace deux contraintes, le coût et l'intelligibilité, et ce seul déplacement suffit à faire du décor l'équilibre spontané. La conséquence pratique est que la gouvernabilité est une propriété *dynamique* : un dispositif récupérable aujourd'hui peut ne plus l'être demain, par simple accumulation de couches et d'automatismes, sans qu'aucune décision explicite n'ait jamais retiré le coupe-circuit. Ce qu'aucun audit ponctuel ne verra, parce qu'il certifie un état et que le problème est une trajectoire.

Le terrain clinique offre un test de cette grille. Tant que ce grounding n'est pas établi, l'exemple reste une hypothèse de travail, pas une preuve.

7. Ce qu'il faut demander, ce qu'on peut mesurer, et où la thèse s'arrête

La conséquence réglementaire est simple à énoncer et coûteuse à tenir. Les textes existants (RGPD article 22 sur la décision automatisée, AI Act article 14 sur le contrôle humain effectif, cadres normatifs comme le NIST AI RMF ou l'ISO/IEC 42001) évoluent déjà vers des notions d'efficacité de la supervision. La critique ne vise donc pas les textes, qui seraient un homme de paille, mais leur opérationnalisation : tant qu'un audit vérifie la présence d'un humain plutôt que la capacité de récupération du dispositif, il certifie un décor.

Reste à expliquer *pourquoi* la présence l'emporte si régulièrement sur la récupération. Ce n'est pas seulement une erreur de conception réglementaire ; c'est une asymétrie de vérifiabilité. La présence d'un humain est observable d'un coup d'œil, auditable sur pièce, juridiquement opposable, et presque gratuite à constater : il suffit d'une signature, d'un horodatage, d'un journal. La capacité de récupération, elle, ne se constate pas, elle se *teste* : il faut la simuler, l'exercer, mesurer ce qu'elle restaure et en combien de temps. Le régulateur, comme l'organisation, optimise spontanément ce qui se vérifie à bas coût. La signature gagne parce qu'elle est cheap to verify, non parce qu'elle est efficace. Toute proposition qui ignore ce coût de vérification restera un vœu.

Il faut aussi anticiper l'effet pervers. Si un régulateur adoptait la capacité de récupération comme indicateur, les organisations optimiseraient l'indicateur sans la chose : des coupe-circuits qui existent sur le papier, des procédures de retour arrière jamais testées. C'est la loi de Goodhart. La parade ne tient pas dans un meilleur indicateur de présence, mais dans un *test d'effet* : la récupération ne se déclare pas, elle s'éprouve, comme on teste une sauvegarde en la restaurant et non en vérifiant qu'elle existe.

Cela autorise un début de métrique, à condition de la traiter comme paramètre d'un test exercé et non comme case à cocher déclarative. Quatre grandeurs sont opposables, et toutes n'ont de sens que mesurées sur un exercice réel :

1. Le *temps de récupération* observé entre la détection et le retour dans l'espace acceptable,
2. La *fraction d'états effectivement restaurés* lors d'exercices, rapportée aux états visés,
3. La *marge avant seuil irréversible*, c'est-à-dire le délai entre le moment où le rattrapage devient possible et celui où il devient vain,
4. La *couverture des scénarios de restauration* réellement joués, et non listés.

Ces chiffres ne valent rien isolés du protocole qui les produit. On revient ainsi au cadrage exigible de toute donnée : ce qu'ils prouvent (une machine de rattrapage a fonctionné, ici, dans ce délai), ce qu'ils ne prouvent pas (qu'elle fonctionnera face à une perturbation non représentée, §4), et pourquoi on les mobilise malgré tout (faute de mieux, une récupération exercée est le seul indice disponible d'une récupération réelle). C'est la même exigence que la thèse maîtresse, appliquée à la mesure : la performance sous perturbation, pas l'inventaire.

Trois limites, enfin, qu'il faut nommer plutôt que dissimuler.

1. La première a été dite : on ne démontre pas la capacité à gérer une perturbation spécifique inconnue, seulement la présence de capacités génériques de récupération éprouvées. L'incertitude sur l'inédit est irréductible (§4).
2. La deuxième : dès qu'on passe du superviseur individuel au collectif (comité de crédit, réunion de concertation pluridisciplinaire, conseil d'administration, comité Théodule…), le mécanisme dominant cesse d'être architectural et devient politique : délibération, coalition, diffusion de responsabilité. C'est un autre objet.
3. La troisième : concevoir un dispositif qui *reste* récupérable quand l'inattendu survient est une question si vaste qu'elle appelle un traitement distinct ; cette note l'a juste mobilisée comme régime décisif, en effleurant le cœur du sujet.

On résume tout ceci d'une phrase. On ne gouverne pas ce qu'on surveille. On gouverne ce qu'on peut récupérer, et seulement aussi longtemps qu'on l'a éprouvé.

8. Conclusion

La question fondamentale n'est pas de savoir si un système est surveillé, ni même de savoir s'il est résilient.

Elle est de savoir si une autorité conserve effectivement la capacité d'en modifier la trajectoire lorsque les hypothèses qui ont présidé à sa conception cessent d'être vraies.

La gouvernabilité n'est pas l'existence de mécanismes de contrôle. Elle s’incarne dans la persistance d'un pouvoir d'action sous perturbation.

Un système cesse d'être gouvernable non lorsqu'il commet une erreur, mais lorsqu'aucune intervention réaliste ne permet plus d'infléchir son évolution.

C'est seulement à cet instant que la supervision devient un décor : une présence observable mais sans capacité démontrée à modifier le comportement du système. Le contrôle subsiste comme attribut organisationnel, juridique ou réglementaire ; il a cessé d'exister comme pouvoir effectif d'orientation. L'autorité demeure formellement présente, mais son action n'altère plus la trajectoire réelle du dispositif.