

**Long read file**

# Human supervision and decision-support systems

---

**October 2024**

[linc.cnil.fr](https://linc.cnil.fr)

**Charlotte Barot**, AI Analyst in the Artificial Intelligence Department

# SUMMARY OF THE LONG READ FILE

<b>HUMAN SUPERVISION, HYBRID DECISIONS: WHAT ARE THE CHALLENGES?</b>	<b>3</b>
<b>DATA ACCUMULATION: THE NEED FOR ANALYSIS</b>	<b>3</b>
<b>AUTOMATING FOR BETTER DECISIONS</b>	<b>4</b>
<b>HUMAN SOVEREIGNTY AND SHARED DECISIONS</b>	<b>4</b>
<b>EFFECTIVENESS OF PROCEDURES</b>	<b>5</b>
<b>BELIEVING OR DOUBTING: THE ISSUE OF CONFIDENCE BIAS IN DECISION-MAKING</b>	<b>6</b>
<b>TOO MUCH TRUST BRINGS RISK: EXCESSIVE APPRECIATION</b>	<b>6</b>
<b>AVERSION AND EXCESSIVE MISTRUST</b>	<b>8</b>
<b>AMPLIFICATION OF BIASES AND HARMFUL INTERACTIONS</b>	<b>8</b>
<b>APPROPRIATE JUDGMENT: INFLUENCES AND CONDITIONS FOR EXERCISE</b>	<b>9</b>
<b>CONCLUSION</b>	<b>12</b>
<b>PREDICTING WITHOUT EXPLAINING, OR WHEN ALGORITHMIC OPACITY MUDDIES THE WATERS</b>	<b>13</b>
<b>PREDICTING IS NOT EXPLAINING</b>	<b>13</b>
<b>TOWARD INTROSPECTIVE SYSTEMS</b>	<b>14</b>
<b>BEHAVIORAL ANALYSIS OF SYSTEMS</b>	<b>14</b>
<b>USER REINFORCEMENT LEARNING</b>	<b>15</b>
<b>OVERALL CONCLUSION</b>	<b>16</b>

# Human supervision, hybrid decisions: what are the challenges?

---

*As decision-support systems become more widespread, their use in critical decision-making contexts raises serious ethical and legal concerns. To mitigate the main risks identified in the field of decision-making, the law requires human intervention or integrated human oversight within the decision-making process, resulting in "hybrid" mechanisms that combine computational power with human judgment. In this series of articles, the LINC explores, based on scientific literature, two key obstacles to the effectiveness of such mechanisms: on the one hand, user trust biases toward the system, and on the other, the opacity of the system's suggestions.*

Human oversight is presented as an essential safeguard to ensure reliable and fair decisions, leading to hybrid decision-making procedures that combine human intervention with the use of algorithms. However, these hybrid procedures, which aim to merge the efficiency of decision-support systems with the human qualities of judgment, can only function if the human decision-maker is able to fully assess the output presented to them, an issue already highlighted by the CNIL in [its 2017 ethical report "How can we ensure that humans remain in control?"](#). In this literature review, the LINC explores the obstacles to implementing hybrid decision-making systems and the potential avenues identified for their improvement.

## Data accumulation: the need for analysis

With the rise of big data, decision-support algorithms (or ADM, automated decision making) have become essential tools for addressing complex problems. These artificial intelligence (AI) systems generate rapid estimates on which decisions must be made, yet without providing objective elements to help assess what the AI proposes. This is particularly relevant in sectors such as healthcare ([Jacobs 2021](#), [Gaube et al. 2021](#), [Beede et al. 2020](#)), finance, content moderation ([Link et al. 2016](#), [Gillespie 2020](#)), or fraud detection.

Decision-support algorithms offer clear benefits when automating simple yet labor-intensive tasks; by doing so, they ease part of the human workload and, in principle, reduce human resource costs. At the individual level, many minor everyday decisions are commonly delegated to algorithms without harmful consequences, such as recommending the fastest route to work or suggesting songs to listen to.

For more complex tasks involving difficult choices, however, trust in these tools can extend beyond delegating the execution of well-understood processes to also handing over part of the judgment itself, turning the system into a guide, or even an “oracle”.

## Automating for better decisions

While delegating everyday decisions may seem reasonable (since they carry little consequence), asking ChatGPT to make an important decision, such as purchasing a property, appears far less so. Such a service is expected to provide recommendations or advice at best, but it would seem unreasonable, for these kinds of issues, to let the algorithm have the final say.

Despite their impressive capabilities, these systems cannot fully substitute for human judgment. For certain types of problems, they prove inadequate for decisions that require not only assessing likelihoods but also considering a broader set of factors, ethical concerns, social context, and long-term consequences of the decision, as in the field of justice.

## Human sovereignty and shared decisions

The use of automated systems raises ethical concerns, one possible response being the integration of human oversight to preserve decision-making autonomy and ensure the relevance of decisions.

Thus, the French Data Protection Act (loi informatique et libertés) excludes, in [Article 47](#), judicial decisions based on automated processing and sets strict limits on the automation of decisions that significantly affect individuals. It requires that whenever a decision has a notable impact on a person’s life, it cannot result solely from a fully automated process and must include human intervention, except in the case of individual administrative decisions.

Similarly, the General Data Protection Regulation (GDPR) sets out the same safeguards in Article 22, refining the scope of possible exceptions: when the data subject has given consent, when automated processing is necessary for the performance of a contract, or when explicitly provided for by law.

## Effectiveness of procedures

- Unfortunately, human intervention does not always improve the outcomes produced by an automated system and can even worsen them. This is due to the fact that decision-makers may adopt rigid attitudes, either by blindly accepting the system's suggestions (known as automation bias or acceptance bias), or by rejecting them on principle (known as aversion bias). Such biases undermine performance and, consequently, threaten the reliability of hybrid mechanisms. The article "[To Believe or to Doubt : Trust Biases in Decision-Making](#)" explores issues related to human decision-makers' trust biases, which affect decision quality, and highlights possible avenues for remediation identified by the research community.
- However, trust issues are not simply errors in judgment : they reveal a deeper difficulty in correctly evaluating the results provided by AI, raising the question of the interpretability of outputs. It is therefore crucial to equip decision-makers with the tools and knowledge they need to determine when to follow algorithm recommendations in order to optimize the overall effectiveness of the system and distribute the burden of decision-making.
- The article "[Predicting without explaining, or when algorithmic opacity clouds the picture](#)" explores the challenges of understanding the results of automated systems, which can be difficult to interpret or question, and details practical ways to facilitate interaction with the proposed outputs and strengthen the reliability of hybrid decision-making procedures.

# Believing or doubting: the issue of confidence bias in decision-making

---

As decision-support systems become increasingly widespread, their use in critical decision-making contexts raises serious ethical and legal challenges. To address the main risks identified in the decision-making process, the law requires human intervention or integrated human oversight within the procedure, resulting in “hybrid” mechanisms that combine computational power with human judgment. In this series of articles, the LINC examines, based on scientific literature, two key obstacles to the effectiveness of such mechanisms: on the one hand, user trust biases toward the system, and on the other, the opacity of the system’s suggestions.

In December 2023, the [Court of Justice of the European Union](#) ruled that the credit scoring tool of the German company SCHUFA, which provided banks with an estimate of a client’s creditworthiness in the form of a trust score, constituted a fully automated decision. In practice, however, the decision to grant credit was formally made by a bank employee responsible for verifying and either applying or rejecting the tool’s proposal, thereby implying human intervention. Yet, since the score was never actually contested by bank employees, who systematically relied on it, effectively turning the suggestion into the credit decision, the Court concluded that such intervention was not meaningful.

To be more than a mere formality, human oversight, as defined in the [European AI Act](#), must make it possible to detect errors, diverge from, or interrupt the system ; in other words, to provide a genuine alternative to the AI system’s output. These abilities, however, rely on psychological dispositions highlighted by the regulation, which specifies that the person responsible for oversight must be aware of potential cognitive biases, such as excessive trust. This bias poses a risk of undermining the quality of the oversight exercised, and consequently, the reliability of the entire mechanism.

Empirical literature does not point to a uniform reaction of individuals toward algorithms but instead identifies two tendencies: either an acceptance heuristic (the appreciation thesis), noted in the European AI Act, or a rejection heuristic (the aversion thesis), despite the errors that both strategies can produce.

## Too much trust brings risk: excessive appreciation

Some studies highlight a tendency among participants to conform to the system’s outputs (e.g. [Jacobs et al. 2021](#), [Green 2019](#), [Yin 2019](#), [Bussone et al. 2015](#), [Kiani et al. 2020](#), [Alberdi et al. 2004](#), [Logg et al. 2019](#), [Robinette et al. 2017](#)). In such cases, human oversight ceases to be

effective, as the decision-maker systematically relies on the system's output, mirroring the logic of the Court of Justice of the European Union's ruling on SCHUFA. Here, it is not the effectiveness of the decision itself that is in question, but rather the effectiveness of human oversight over algorithmic decision-making, and its ability to reject potential system errors or anomalies.

In a study conducted in psychiatry ([Jacobs et al. 2021](#)), researchers asked a cohort of clinicians to make a series of decisions about a fictitious patient. For each patient, the clinician had to decide on a treatment, either with a recommendation from a machine learning system accompanied by a brief justification of the suggested choice, or entirely without any recommendation (a completely independent decision).

The study shows that, on the one hand, the performance of the groups with and without access to the machine learning system's recommendations is virtually the same, and that both groups make poorer decisions than the system alone. On the other hand, when the system produces an incorrect decision (where the error is defined as a position diverging from that of a panel of psychiatry experts), performance drops compared to the control group (without access to the machine learning system) that is, humans tend to conform to the algorithm's decision in such cases.

Finally, the authors observed an effect of familiarity with the tool : clinicians who were more familiar with the system were, on average, less likely to follow a machine learning system's recommendation, regardless of its accuracy, compared to clinicians who were less familiar with such systems.

The observation that the level of professional expertise plays a role in decision-making is corroborated by a study by [Gaube and al. 2021](#) in which two groups of doctors with different levels of medical expertise were asked to produce a diagnosis and rate a recommendation displayed as coming from either an algorithm or a human, when in fact all the recommendations came from a human.

Doctors in the most expert group tend to rate recommendations less highly when they are indicated as coming from an AI system. However, the quality of their diagnosis is influenced by the quality of the advice received, regardless of its source (AI or human).

This result suggests that the influential effects of the system's advice could be a simple artifact of not having been able to form a decision before the confrontation, and not necessarily due to an attitude of deference toward the algorithm. Indeed, the study shows that participants generally tended to follow the advice given, whether it came from a human or a machine.

Absence of deference to the system by experts has also been observed in other studies ([Logg et al. 2020](#), [Povyakalo et al. 2013](#)). In the latter study, assessment was also modulated by disagreement. Advice that contradicted the participant's prior opinion had less impact but did not completely cancel out the effect of trust in the system. The assessment of algorithms decreased (but did not disappear) when their advice was contrary to their own judgment. The authors conclude that it is on these points of disagreement that people are most likely to

improve their accuracy. These critical cases of confrontation are therefore interesting for decision-making.

## Aversion and excessive mistrust

Some studies suggest excessive mistrust of the system, preventing participants from making informed decisions (e.g. [Dietvorst et al. 2015](#), [Longoni et al. 2019](#), [Dzindolet et al. 2002](#), [Lim et O'Connor 1995](#), [Yeomans et al. 2019](#), [Promberger et Baron 2006](#)).

In several forecasting tasks where they had to choose between an algorithmic prediction and a human prediction regarding the likelihood of a song's success ([Dietvorst et al. 2015](#)), when they saw the system working, and sometimes getting it wrong, participants tended to reject its predictions in favor of human advice, despite the higher error rate of human predictions (up to twice that of the algorithm).

In short, humans are less forgiving of algorithms' mistakes and tend to generalize their overall performance based on these harmful examples. This tendency persists even after observing that the system's performance exceeds that of humans on average. A natural explanation is that mistrust of algorithms exists even before observing them, and that observing mistakes reinforces this preconception.

There would therefore appear to be an anchoring bias at work: once users have formed an opinion about an AI system, it is very difficult for them to change their minds, even in the face of contradictory evidence. This is supported by the fact that aversion to algorithms is not observed in purely deterministic tasks such as logical calculations or memory tasks, where humans are notoriously more fallible than algorithms.

## Amplification of biases and harmful interactions

Humans are not free from bias and do not have infallible judgment, which leads to errors that, all other things being equal, are added to those of the systems or amplify them.

**THE COMPAS CASE**

The [COMPAS](#) (Correctional Offender Management Profiling for Alternative Sanctions) system is a prediction tool used in the criminal justice system in the United States. Developed by Equivant (formerly Northpointe), this system assesses the risk that an offender will commit further offenses or fail to appear at a future hearing. COMPAS uses a set of questions about offenders' criminal history, behavior, and personal characteristics (social relationships, demographic data: age, gender, ethnicity, etc.) to calculate risk scores, including the risk of violent recidivism. Judges often use COMPAS scores to decide whether a defendant can be released pending trial. A 2016 investigation by ProPublica revealed that COMPAS had discriminatory biases, overestimating the risk of recidivism in certain individuals and leading to harsher sentences, without this being justified by criteria other than ethnicity.

To determine whether these biases stem solely from the algorithm [Green et al. 2019](#) replicated the COMPAS system under laboratory conditions. The experimental task consisted of assessing the risk of recidivism of an individual, following the same logic as the original algorithm. Each individual's description was accompanied by a recidivism risk score expressed as a percentage, and the participant had to indicate their own risk assessment. The researchers found that when an algorithmic suggestion was provided, participants made poorer decisions than the algorithm itself and were unable to evaluate either their own performance or that of the algorithm. They also emphasized that decision-makers were clearly biased against certain profiles, which led them to amplify the algorithm's bias. Human biases, combined with high trust in the algorithm, can therefore result in not only accepting a biased outcome but also reinforcing its skew.

On this topic, see also the interviews of Angèle Christin (« [Les méthodes ethnographiques nuancent l'idée d'une justice prédictive et entièrement automatisée](#) ») and of Philippe Besse (« [Les décisions algorithmiques ne sont pas plus objectives que les décisions humaines](#) ») on the website of the LINC.

## Appropriate judgment: influences and conditions for exercise

The observations outlined above highlight the need to pay particular attention to the conditions that enable the user to be in the best possible psychological state for exercising sound judgment. This concerns not only the user's own skills and the task to be performed, but also a range of external factors related to the work environment and the context in which the decision is made. Among these, one can note :

The decision-making environment:

- **The cost of an error or being involved in a high-risk situation** (e.g. Robinette and al. 2017). The risks assumed by the decision-maker, given the consequences for the affected individual, influence their judgment, as following or diverging from a system's recommendation can lead to different outcomes in case of an error. In particular, **the assignment of responsibility to the decision-maker** constitutes a cost borne by that individual. If responsibility for a decision is attributed to the algorithm, it may seem costly for the human involved to diverge from the algorithm's decision, which encourages them to accept the algorithm's recommendations.
- **The time allocated for decision-making** (e.g. [Robinette et al. 2017](#)). Very short deliberation periods lead to decisions that closely follow the algorithm's suggestions, due to a reflexive effort to conserve cognitive resources.

Domain expertise:

- **The level of expertise in the given task** (e.g. [Jacobs et al. 2021](#), [Logg et al. 2019](#), [Povyakalo et al. 2013](#)). Experts tend to rely less on the algorithm than novices, which is advantageous for the ability to diverge from the algorithm's proposed decision, but it can also lead the expert toward potentially excessive confidence. Conversely, less experienced individuals are more likely to accept the system's recommendation more frequently.
- **Doubt regarding the decision to be made**. Cases in which the decision-maker is uncertain or experiences a high level of doubt are likely to be those in which the system's suggestion has the greatest influence. This may occur in difficult situations due to an overwhelming number of options or because the case at hand is unique.

Relationship with automated systems:

- **Familiarity with algorithms and automated systems** (e.g. [Jacobs et al. 2021](#)). Individuals who are less familiar with these systems tend, depending on the situation, either to exhibit greater trust or, conversely, excessive distrust, but their disposition is rarely neutral.
- **Prior trust in the system**, i.e. preconceived notions about the algorithm's reliability, which are very difficult to overcome, especially if they are based on observations that the system has made errors. (e.g. [Robinette et al. 2017](#), [Dietvorst et al. 2015](#), [Prahl et al. 2017](#)).
- **Congruence, or the intuitive agreement with the algorithm's output** (e.g. [Logg et al. 2019](#)). This seemingly trivial factor can strongly impact the human ability to question the output in cases of excessive distrust toward the algorithm.

Finally, the overall configuration in which the interaction between the decision-maker and the decision-support system takes place:

- **Anticipated suggestion.** The system provides a recommendation even before the human makes a decision. The human then decides whether to accept the recommendation or to diverge and propose an alternative. As noted earlier, this configuration can encourage anchoring effects, making it more difficult for the human to diverge.
- **Doubt resolution.** The system does not initiate any action but flags a situation in which it detects a potential risk. In this case, human intervention occurs only when an employee decides whether to validate the alert. Often used in remote monitoring, this configuration reduces the need for continuous human involvement, as personnel intervene only in certain situations, those indicating danger or requiring action. It requires that all potentially hazardous situations be accurately identified by the system.
- **Alternative suggestion.** The system provides additional information to support a human decision that has already been considered. This approach helps prevent overreliance on the algorithm, but it can also reduce the effectiveness of its use, as the human operator has already made a decision and may find it “costly” to change it. However, in cases where the human is uncertain about their decision, this configuration appears to be particularly relevant.
- **Use of a choice algorithm.** A choice algorithm determines whether the decision is returned to the human or handled solely by the system (a single, uncontested decision). This approach, proposed by [Mozannar et al. 2023](#)
- [Mozannar et al. 2023](#) helps avoid human trust biases, but it relies on another automated system, which does not meet the requirements of the GDPR and the AI Regulation, unless multiple layers of human intervention are implemented.
- **Index in a bundle.** Independently, a human and an AI make a decision on a problem; a third party, such as a panel or another expert, resolves any disagreements, ideally in a blind evaluation.

The «**confrontation**» options, namely options 3 and 5, where the user has the opportunity to form an independent judgment, appear best suited to allow the person not to simply “submit” to the system’s decisions, and thus to exercise their judgment. However, for these configurations to be effective, the human must be willing to genuinely compare their decision with the system’s suggestion, without necessarily trying to confirm their own initial hypothesis. While these two configurations are therefore relevant in the context of human oversight, the cases that are likely to be truly beneficial are those where the agent’s opinion is not already strongly fixed; otherwise, there is a risk of simply rejecting the system’s decision.

## Conclusion

Even with perfect neutrality, infallible judgment, and ideal working conditions, a fundamental obstacle remains: the very ability to decipher and make sense of the system's suggestion. It is this intrinsic difficulty that we explore in the article "[Predicting without explaining, or when algorithmic opacity muddies the waters](#)".

# Predicting without explaining, or when algorithmic opacity muddies the waters

---

**As decision-support systems become more widespread, their use in critical decision-making contexts raises profound ethical and legal concerns. To counter the main risks identified in the field of decision-making, legislation requires human intervention or integrated human oversight within the decision-making process, leading to “hybrid” mechanisms that combine computational power with human judgment. In this series of articles, the LINC examines, based on scientific literature, two key obstacles to the effectiveness of such mechanisms: user trust biases toward the system and the opacity of the system’s suggestions.**

The second article in this series illustrated the trust biases at play in assisted decision-making and how they act as an obstacle to the free exercise of human judgment in a hybrid system. While these biases are linked to specific abilities and contextual factors inherent to decision-making, they also reflect an intrinsic difficulty in how the system itself operates: the ability to interpret its outputs. Indeed, when the decision-maker cannot evaluate the system’s suggestion, they are left either to rely on their own intuition or to adopt a heuristic of doubt or trust. Even the minimum condition required for meaningful human oversight, namely, rejecting obvious errors, is not always realistic in contexts where certain outputs are difficult to assess. On the one hand, absurd responses may sometimes appear plausible, for example, in text-generating AI systems that insert an invented name or fact into an otherwise coherent passage. On the other hand, the very format of the output can make it resistant to evaluation: assessing the validity of a numerical score in order to propose an alternative would, in most cases, require reproducing the inference that produced the score in the first place

## Predicting is not explaining

In 1999, in a behavioral study, [Goodwin et Fildes](#) found that when decision-makers were presented with trend predictions in the field of marketing, they tended at best to ignore reliable predictions, and at worst to degrade them by attempting to modify them. In other words, they showed a bias of distrust toward the algorithm. The authors noted, however, that the outputs were difficult for decision-makers to evaluate, since their format, given as a score or percentage, was not easily contestable. When users displayed such maladaptive attitudes, it was because they were unable to interpret the score provided and therefore preferred to ignore it most of the time. Yet when they did attempt to propose an alternative, they failed to perform better than the system.

As this article shows, those tasked with evaluating system outputs are ultimately responsible for two subtasks: understanding and evaluating. On the one hand, the user must make sense

of the system's suggestion, for instance, if the output is a score, they must understand the scale on which it is expressed and the thresholds considered critical. The user must be able to interpret the overall message conveyed by the system (the number) within its context (the scale and thresholds).

Next, the user must evaluate it, that is, make a judgment about its relevance, either accepting it or rejecting it in favor of their own opinion or a corrected version. In the context of machine learning systems, however, evaluating outputs is not straightforward, since these models operate as black boxes. When the inference process that generated the system's suggestion cannot be retraced or unpacked, alternative ways of interpreting the outputs must still be found.

## Toward introspective systems

One option could be to ask the system itself, prompting it to produce justifications that would both help explain its output and allow an assessment of its own reliability. Unfortunately, the justifications and confidence scores accompanying the responses are not always reliable, even when the outputs themselves are correct.

Thus, [Jin et al. 2024](#), analyzing the performance of a model tasked with solving clinical cases based on medical imaging, observed the model's limited ability to justify its answers. The model was tested using prompts structured in three parts: first, it had to describe the provided medical image, then recall relevant medical information to address the question, then produce a medical reasoning process, and finally choose a diagnosis from a set of options. While the model demonstrated high accuracy -sometimes even surpassing that of physicians - in its final diagnoses, it struggled significantly with understanding the medical images. This often led it to produce flawed reasoning to support an otherwise correct diagnosis, thereby generating misleading justifications.

These limitations make its use in clinical settings still premature, as such weaknesses threaten its potential integration into medical practice. The risk of introducing misleading justifications for instance, in a case where the human expert does not have access to the image, or relies on the system's misinterpretation of it would be to mislead the decision-maker, possibly even leading them to reject an otherwise correct final suggestion.

## Behavioral analysis of systems

In sum, systems do not always possess strong introspective abilities: they are not always able to analyze their own behavior, whether good or bad. However, one can still find guidance without attempting to open the black box, by relying instead on a behavioral analysis of the model.

It is in this context that a system developer plays a crucial role in the proper integration of the model into a domain-specific expertise process. The developer can provide several contextual elements that offer a clearer understanding of the conditions under which the model was “trained,” helping to interpret its behavior:

- the context in which the algorithm was designed,
- its known limitations,
- tests conducted prior to deployment,
- tasks on which it typically performs less well, etc.

Information about the training data, the system’s behavior in real-world situations, and the error margins observed during testing also helps to shed light on its outputs.

## User reinforcement learning

Exploratory research delves deeply into this notion of behavioral analysis of models by providing users with “training” models of human decision-makers to familiarize them with the behavior of the system in use ([Lian et Tan 2019](#), [Suresh et al. 2021](#), [Wortman Vaughan et Wallach 2021](#)). The goal is to teach users, through a series of trials, to become acquainted with the system’s behavior, enabling them to know when to follow its suggestions, when to reject them, and, in the latter case, when to investigate the problem more thoroughly.

In their experimental setup, [Mozannar et al. 2022](#) explore optimizing collaboration between humans and artificial intelligence systems on tasks involving question answering based on text passages (using the [HotPotQA](#) dataset). The article proposes a method to help users collaborate with different AI models: by the end of the training, they should be able to decide when it is preferable to delegate the answer to the model and when they should intervene themselves.

This method draws on educational research emphasizing the importance of feedback in learning. It is based on the principle of specific examples, which are prototypical cases designed to illustrate situations in which the algorithm is reliable and those in which it is not. The examples are selected to represent different scenarios: some in which the algorithm has a high level of confidence and makes a correct prediction, others in which confidence is high but the prediction is incorrect, and cases where the confidence level is uncertain, regardless of whether the prediction is correct or not.

The goal is to improve the “mental model” that humans form of the algorithm’s capabilities, that is, to help them understand the cases in which it is likely to make errors, including errors in its own confidence estimates. This learning process enables users to better recognize situations where they can trust the algorithm and those where, conversely, it is necessary to verify the results more carefully.

Experiments show that users trained with this method are more effective at deciding when to delegate decisions to the classifier, enhancing collaboration between decision systems and humans and reducing judgment errors.

## Overall conclusion

Scientific literature shows that implementing hybrid decision-making systems involves two types of challenges: first, enabling the decision-maker to exercise judgment that is, in principle, informed and impartial, which depends on exogenous conditions of the decision-making process; and second, allowing the decision-maker to correctly interpret the system's outputs, which depends on intrinsic conditions of system readability. Ultimately, the decision-maker's trust attitudes merely reflect these underlying conditions. These two types of obstacles indicate that responsibility for hybrid decisions must be shared between the system deployer, who bears the operational risk, and the system designer, who is responsible for the system's proper functioning and for providing the tools needed to learn how to use it effectively.

Indeed, if human intervention requires a significant degree of discretion, the risk is that individual responsibility in decision-making increases proportionally: the greater the freedom, the greater the associated costs borne by the decision-maker. Beyond the decision procedure itself, it is therefore necessary to broaden the perspective and consider these new procedures within the full work context, in order to best integrate machine suggestions into human decision-making.